THE MISLED MIND

by

RICHARD PAUL FARRY

A thesis submitted to the University of Birmingham for the degree of

DOCTOR OF PHILOSOPHY

Department of Philosophy
School of Philosophy, Theology, and Religion
College of Arts and Law
University of Birmingham
June 2025

Dedication

It's been a long road, getting from there to here.

This is my first ever acknowledgements page, please bear with me.

I wouldn't have been able to do this if it wasn't for the initial support from my boss. He is too nice to deserve to be called a 'boss', but here we are.

This is all my own work. But I would never have been able to do to the work, if it wasn't for the support and sacrifices of my family. It's been a long road, much longer than we had planned.

I also would not have had the ability to do the work, if not for the support of my supervisors. They've put up with me, also for much longer than we had planned.

I think that sometimes my supervisors think that some of my views are crazy. But from talking to them individually, they seem to disagree over which views are the craziest, so I hope it all balances out.

Thank you everyone for supporting me and putting up with me. You are all amazing.

I hear in Finland when you get a PhD they give you a sword. I probably don't deserve one, but I'm not going to let that stop me.

Acknowledgements

The fees for this PhD were paid for by QinetiQ plc.

Contents

THESIS INTRODUCTION: THE MISLED MIND

1 Introduction

In this dissertation, I shall defend the overarching thesis that the way we experience our own minds misleads us as to the actual nature of our minds.

This dissertation is made up of three related papers that support the overarching theme. What links each of the topics is that they champion a counter-intuitive view about how our minds work. Simply put, the way our minds seem to us—the way we experience them—misleads us about the reality of our minds. We are misled in three major ways: that we are distinctively self-knowing minds, that there is something special about our experiences which might not be explainable by science, and that we are minds confined within our bodies. I take on each of these intuitive but mistaken views in the three papers, seeking to reverse or subvert them.

1.1 The Core Problem

In seeking to reverse or subvert our intuitive views of our minds I face a challenge. Counterintuitive views about our own minds are particularly difficult to accept. There are two broad reasons for this.

The first reason is that our experience of our own minds comes from a rather limited and distorting perspective, but we seem predisposed to consider what we get from that perspective to give us a full picture. Part of why we have a limited perspective is

that we are blind to much of the workings of our minds, of how our thoughts arise. Much of conscious thought arises from the darkness of our unconscious. I do not know how I get the answer to a relatively easy and familiar multiplication; I just get the answer. Maybe my brain has recalled the answer, or maybe it has calculated the answer, but from my perspective I cannot tell which strategy was used. I often do not know exactly what I will say next, but the words come. What our minds are and how they work is like a black box to us. We are highly reliant on the conscious products of unconscious mental activity as a base for understanding our minds. This is not true of all our conscious mental activity[1]; our deliberate and effortful conscious mental activities seem to be something we are aware of and can direct. However, it is broadly true of much of which we are consciously aware. We get the 'output' of mental activity, without an awareness of how it came about. Because of this, we do not have any (or very limited) first-person basis to question what we are conscious of. And it is these outputs, and their appearances, that form the basis of common-sense folk-psychological views and intuitions about the mind. This is like thinking we know what an iceberg is like based only on what we see above the surface of the water.

The second reason is that our common-sense intuitive sense of the mind is difficult to overcome, even in the face of competing evidence. Consider how we perceive tables as solid objects, while in reality, they are mostly empty space (Eddington, 1927). This is akin to what Sellars called the manifest and scientific image (Sellars, 1962). The idea that a solid table is largely empty space is, in everyday terms, ludicrous and conflicts

[1] Given the nature of my project here, I make this qualification tentatively even though I take it to be true.

with common sense. Yet, we accept these two seemingly incompatible images of the table as the way things are. We never outright reject the manifest image of the table as a completely solid object or reject our everyday folk-physics understanding of what objects are. Perhaps we never will, because it would be impractical and perhaps counterproductive to do so. But in not rejecting the manifest image of what tables really are, we also do not reject the scientific image, holding both images of the table to be true. We should adopt the same mindset when considering our minds, but it is challenging to do so. In some respects, the way things seem to us is somewhat like a visual illusion, in that we cannot shake it off, even when presented with evidence of its falsity.

That our knowledge of our own minds is fallible, and that we may be substantially wrong about them, is not a new thesis. In many ways it is quite a mainstream view; for example, the idea of our unconscious minds secretly influencing what we do and say is widespread. There are likely few philosophers of mind who would accept that how our minds seem to us gives a complete and accurate picture of how our minds really are.

And yet, my worry is that—quite naturally and understandably—we still place too much emphasis on how things seem to us at the expense of fully exploring or endorsing counterintuitive views about the mind. There is much wisdom in relying on common sense as an anchor point in our philosophical theorising, but also some folly in holding onto it too tightly. In each of the following papers I take a topic in the philosophical literature that relates to our minds being rather different to how they seem to us. For each topic I argue that not only is the view championed in them

correct, but the counterintuitive aspects of mind that they articulate may be more pervasive than supposed.

Each paper presents a view about our minds that is counterintuitive and somewhat odd, but their apparent oddity arises from our limited and parochial perspective on our minds—the 'view from the inside'—which means we cannot help but find the view counterintuitive. Whilst the views explored in the three papers may seem counterintuitive, they are entirely plausible, even somewhat compelling.

2 Background

We have a sense of what things in the world are like, what they are. But as we learn more about how the world works, often our naïve views of the world prove to be wrong. While collectively we have made progress in understanding the mind and the brain, along the journey there have been some surprising and counterintuitive discoveries. The idea that the way our minds seem to us is misleading is not radical, but some of the discoveries about our minds have been. While we might expect to get quite a bit wrong about various matters of the external world, it seems astonishing that we could be wrong about our own minds in any meaningful way. Yet, people in the past have held completely different views about their minds to what we hold now. Thus, we should be open to the idea that we might be misled about our own minds.

An early radical shift in our thinking about the mind involves the role of the brain. While the consensus view is that the brain is the primary seat of the mind and origin of our thoughts, this has not always been the case. In the fourth century BCE, Aristotle

considered the brain's purpose to be to cool the blood (Longo 1996; Rábano 2018). To our contemporary intuitions, this seems strange almost to the point of absurdity. It is hard to shake the idea that 'I' reside behind my viewpoint of the world. Yet clearly, Aristotle was operating on a different set of intuitions about the mind. It was not until centuries later, that through his surgical investigations, Galen (130CE – 210CE) came to think that the brain was home to our rational part, and the location where the information from our senses came together (Freeman 1994; Singer 2021).

Views of the mind and the self have changed remarkably as well. In early tales, such as Homer's *Iliad*, there is a sense that we are moved by forces of which we're not consciously aware. For Jaynes (1976, pp.67-83; Moore 2010), the characters in the *Iliad* generally lack consciousness, and the voices in their heads are taken to be the commands of external gods, rather than an inner voice. Later writing in the Ancient Greek world seems to indicate that they had no distinct view of the self that we would find familiar (Gill 2017). Instead, various psychological activities were considered to be separate 'psychic entities' within themselves, rather than part of a cohesive whole (Sullivan 1995, pp.14-15). Lyons (1986, p.1) suggests that the Ancient Greeks did not even have a concept of introspection, and the first early sign of such a concept arguably occurs in Aristotle's *De Anima*. It was not until much later that literature showed concepts of the self that would be more recognisable to us now.

It has not been a straight journey between the ancients and us when it comes to how we think of the self. There was, potentially, a kind of self-awareness dark age. Nørretranders (1999, pp.319-323) argues that self-awareness in Europe dwindled

between 500 and 1,000 CE, corresponding (at least in part) to the decline in the availability of mirrors in that period and subsequently rising with their improved availability. The idea here is that seeing our reflections gives us a different perspective of ourselves in comparison to and distinct from others which contributes to a sense of self. While we might be familiar with the idea that different metaphors of the mind might give us different ways to think and talk about the mind, it seems rather strange that the presence of mirrors could contribute to our sense of selfhood. Again, we should be cautious in accepting what seems obvious about our minds based on our current intuitions and concepts.

Closer to our own times, the work of Descartes has dominated and shaped how we view our minds for centuries, giving us a view of ourselves that Ryle (2000) referred to as 'The Ghost in the Machine': that we are a distinct observer of what Dennett (1991) dubbed 'The Cartesian Theatre'. Under this Cartesian view, we are an immaterial substance that (non-spatially) inhabits or controls our physical bodies, a kind of homunculus viewing the presentation of experiences. This is a compelling and difficult to shift metaphor for how the mind works, which can often pop up in everyday discourse and entertainment media (e.g. Pixar's *Inside Out* movies).

This Cartesian thinking about the mind has come under pressure from a variety of directions[2], including our growing awareness and understanding of the unconscious mind: the radical idea that there are a whole host of activities going on in our brains, of which we are unaware, that give rise to or underpin our conscious thoughts. While

[2] Including the difficulties faced by Cartesian dualism.

Freud developed the idea of the unconsciousness and helped to popularise the concept in the 20th century, it had long been a topic of interest and inquiry to philosophers. For example, Schopenhauer (1851/2004 s.17) wrote about how we can contemplate a problem, and a few days later the answer will occur to us "...entirely of its own accord; the operation which has produced it, however, remains as much a mystery [..] as that of an adding machine". Nietzsche (1882/2001 s.354) thought we had both conscious and unconscious thoughts, and William James, a founder of modern psychology, is thought to have been a supporter of the idea of unconscious processes (Weinberger 2000).

Pressure on the Cartesian view has arisen from empirical findings and theoretical concerns about the reliability of the access we have to our own minds, and the extent to which that access might distort or mislead us as to the reality of our minds. Early in the 20th century, the scientific use of introspection—of turning our attention inwards—in psychological study was a serious endeavour. Introspection is a key way through which we experience our minds, so is an important part of how we come to understand them. But there were serious doubts as to the quality of the outputs of introspection and its use in psychology. Some of the methods at the time required strenuous training, involving, in some instances, thousands of trials of practice, or the use of a 1,600-page training manual (Schwitzgebel 2010) to meet the requirements of the introspectionist scientists of the day.

The study of introspection came to be taken less seriously or to be outright rejected with the rise of philosophical and psychological behaviourism. This was concerned with

external behaviour and neglected or denied inner behaviours and mental states (Graham 2023 and Hauser 2024). By most accounts, this state of affairs largely persisted until the 'cognitive revolution' of the 1950s and 1960s, at which point the study of the mind and of mental states became an active area of interest in experimental psychology (Miller 2003).

The cognitive revolution reintroduced internal states and processes as key to understanding the workings of the mind (Kosslyn *et al*. 1995; Miller 2003; and Thagard 2023). This supported the exploration of capacities and concepts such as memory, attention, mental models, and visual processing. Crucially, these involved internal states and processes which were not posited simply based on our first-person awareness of them. Indeed, in some cases we do not have first-person awareness of them. This inclusion of states and processes that are not first-person observable broke the link between our commonsense picture and our scientific picture of the mind. From the cognitive revolution onwards, we can no longer assume we are standing on firm ground when we rely only on how our minds seem to us, to explain how our minds work.

The acceptance of introspection as a worthwhile route to scientific knowledge may have taken some time to shake off the effects of the behaviourist turn, with continuing resistance in some quarters as to its respectable use in serious science due to its subjectivity and apparent issues of reliability. However, as Spener (2024, pp.3-7) indicates, many of these apparent issues arise from a lack of clarity in the various literatures of what is meant by the term of introspection—with a conflation of the

capacity of introspection with the use of that faculty—and the various different targets of introspection. Introspection remains an important route to understanding the mind, though as Dennett (2013, pp.341-346) advises, we should perhaps focus on the objective third-party study of introspective reports. With such an approach we can take seriously people's self-reports about what experiences are like for them, and what they believe to be true about their mental lives, but we do not need to make the mistake of assuming that those beliefs are true. Introspection is a useful and powerful means for us to explore and understand our minds, but we should not be misled by it or rely solely on it as a guide to our minds. Even the brief history I have outlined here shows that we should be somewhat mindful as to how much we read into how our minds seem to us based on our intuitions and introspection.

What follows is an outline of three intuitive views of the mind, with the counterintuitive position that I argue for.

3 We Are Not Distinctively Self-Knowing Minds

We have mental states like beliefs, intentions, and preferences. These states are called propositional attitudes. The typical view of how we know our propositional attitudes is that we just kind of do, we have a sort of direct, reliable, and immediate access to them. If I consider whether I believe there is a monster under my bed, I easily know what the correct answer is[3].

[3] 'No'.

The counterintuitive view that I defend in my paper is that a portion of what we say and think about ourselves is fabricated; that many of our declared propositional attitudes are just made-up. Made-up *well* in many cases, such that these fabrications go unnoticed or are glossed over, but nevertheless made-up and delivered with great confidence in their veracity. These fabrications are called confabulations.

According to this view, rather than directly accessing our propositional attitudes like our beliefs and intentions, in some cases we interpret them based on theorising and available evidence (Carruthers 2013). That is, we figure out what our propositional attitudes are. Imagine that our mental states were in individual boxes, if this were the case then the current intuitive view would have us observing or retrieving the contents of a box when we want to know what one of our propositional attitudes is. In contrast, under the interpretative view I'm defending, it would be more like working out what was inside the box via detective work. As a result, we do not know as much about our own mental states as we think, and we confabulate as a regular everyday occurrence, generally without anyone noticing.

In my first paper—'In Defence of the Interpretative Sensory-Access Theory of Self Knowledge'—I defend this counterintuitive view from objections which deny that we are confabulating when we report our propositional attitudes. The first objection accepts that sometimes our propositional attitudes come about due to confabulation, but we subsequently access them directly (i.e. we do not interpret them). My response to this line of argument is that the objection does not hold in those cases where we are reporting what our propositional attitudes were before or at the time of questioning.

The second objection argues that what appears to be confabulations are in fact our propositional attitudes being influenced by perceptual cues, which are subsequently accessed successfully. My response is to point out that if we were to accept that position, it would make us far more mercurial beings than we are.

4 There Is Not Something Special About Our Experiences Which Might Not be Explainable by Science

We see the rich *redness* of an apple, feel the *painfulness* when we stub our toe. On the face of it, the *redness* and *painfulness* experiences do not seem like something a brain could bring about. A brain is after all, despite its complexity, a big hunk of neurons signalling at each other, which does not seem like the kind of thing which could lead to experiences.

The qualitative aspects of experiences—such as *redness* and *painfulness*—are often called phenomenal properties. The counterintuitive view of illusionism, which I argue for, holds that phenomenal properties do not exist. There is no such thing as *redness* or *painfulness*; they are misrepresentations—illusions. They only seem to exist, much like rainbows only seem to exist as real objects in the sky. According to illusionism, whatever properties give rise to our experiences of things like *redness* or *painfulness* they are not separate or distinct from the representational or functional properties involved in our experiences (Frankish 2023, p.5).

In my second paper—'Useful but Not Accurate'—I provide an argument in support of illusionism. I argue that as an evolved cognitive organism we should expect and find how our conscious experiences appear to us (e.g. *redness* and *painfulness*) is a kind of useful fiction or model. As such, we should not expect these representations (and often do not find them) to be accurate in terms of what they represent. While these experiences are useful, they are also deeply misleading. The character of conscious experiences shape how we view the world and ourselves. The way things seem to us is not how they really are. We may never shake off this perspective, and we may never want to, but we should accept it as an appearance and be more accepting of the apparent gap between appearance and reality.

5 We Are Not Minds Confined to Our Bodies

In my last paper—'Where Is My Mind?'—I consider the boundaries of our minds. Our natural inclination is to see our minds as constrained within the 'boundaries of skin and skull' (Clark and Chalmers 1998), in what Ross and Ladyman (2010) refer to as 'the container view'. However, I argue that processes in general often outrun the physical borders of objects, and that we should accept this and not be led by our intuitions and common sense into making an exception for cognitive processes. The result is that we should recognise the reality of 'cognitive bloat', i.e. that our environment is filled with things that can constitute part of our cognitive processes, and that such constitution is in fact a frequent, everyday occurrence.

The consequence of my view is that various objects that we make use of are part of our cognitive apparatus, helping us to think and to remember. As part of our cognition, they're part of our minds, at least temporarily. As such we should consider and treat them differently compared to regular objects, because there is added importance in safeguarding these external parts of our minds.

6 Summary

There is substantial heterogeneity across the phenomena championed in this dissertation. There is also heterogeneity in what gives rise to these misleading experiences of the mind. In the case of self-knowledge, for example, it seems likely we are misled by the nature of our cognition, by the seeming immediacy and reliability of our self-reports. Whereas, in contrast, with the case of phenomenal properties, while we may have pre-theoretic intuitions—e.g. many lay people would say a robot cannot experience *pain*—the notion that there is something potentially anomalous about the brain bringing about *redness* or *painfulness* follows on from a set of philosophical commitments.

Overall, in this dissertation, I argue the case that our minds are not how they intuitively appear to us or how we naively or commonly tend to think about them. Further, the counter-intuitive aspects of mind, championed in this dissertation, are more pervasive than is normally supposed. Our basic perceptions and conscious experiences of the world are shaped to be useful rather than accurate; many of our self-reports about our judgements and propositional attitudes are *post hoc* fabrications; and our cognition makes use of, or is even dependent upon, external objects in our environment, such

that our minds do not only reside within our heads. Our experience of our minds misleads us as to their nature, which is difficult and counterintuitive to accept, but if successful the following three papers will at least raise a few doubts, if not persuade you to hold a different perspective.

7 References

Carruthers, P. (2013) *The Opacity of Mind*. Oxford University Press.

Chalmers, D. (1995) 'Facing Up to the Problem of Consciousness', *Journal of Consciousness Studies*, 2, 3, pp.200-219.

Chappell, S. and Verde, F. (2025) 'Plato on Knowledge in the *Theaetetus*', in in Zalta, E. and Nodelman, U. (eds.) *The Stanford Encyclopedia of Philosophy* (Spring 2025 Edition). Available at: https://plato.stanford.edu/entries/plato-theaetetus/ accessed 24/06/2025.

Dennett, D. (1991) *Consciousness Explained*. Penguin.

Dennett, D. (2013) *Intuition Pumps and Other Tools for Thinking*. Allen Lane.

Eddington, A. (1927) 'Gifford Lectures: Introduction'. Available at: https://mathshistory.st-andrews.ac.uk/Extras/Eddington_Gifford/ accessed 14/09/2024.

Frankish, K. (2023) 'What is Illusionism?', *Klēsis Revue Philosophique*, 55, pp.1-17.

Freeman, F. (1994) 'Galen's Ideas on Neurological Function', *Journal of the History of the Neurosciences*, Volume 3, Issue 4, pp.263-27.

Gill, C. (2017) 'The Self in Greek Literature', *Oxford Classical Dictionary*. Available at: https://oxfordre.com/classics/display/10.1093/acrefore/9780199381135.001.0001/acrefore-9780199381135-e-8109 accessed 20/02/2025.

Graham, G. (2023) 'Behaviourism', in Zalta, E. and Nodelman, U. (eds.) *The Stanford Encyclopedia of Philosophy* (Spring 2023 Edition). Available at: https://plato.stanford.edu/entries/behaviorism/ accessed 14/09/2024

Hauser, L. (2024) 'Behaviourism', *Internet Encyclopaedia of Philosophy*. Available at: https://iep.utm.edu/behaviorism/ accessed 14/09/2024

Jaynes, J. (1976) *The Origin of Consciousness in the Breakdown of the Bicameral Mind*. Houghton Mifflin Company: Boston.

Kosslyn, S., Berhmann, M. and Jeannerod, M. (1995) 'The Cognitive Neuroscience of Mental Imagery', *Neuropsychologia*, Vol.33, No.11, pp.1335-1344.

Longo, O. (1996) 'Hot heads and cold brains. Aristotle, Galen and the "radiator theory", *Physis; rivista internazionale di storia della scienza*, 33(1-3), pp.259-266.

Lyons, W. (1986) *The Disappearance of Introspection*. The MIT Press.

Miller, G. (2003) 'The Cognitive Revolution: A Historical Perspective', *Trends in Cognitive Sciences*, Vol.7, No.3, pp.141-144.

Moore, J. (2010) "'They Were Noble Automatons Who Knew Not What They Did:" Volition in Jaynes' The Origin of Consciousness in the Breakdown of the Bicameral Mind', *Frontiers in Psychology*, Vol.20, No.12, pp.1-4.

Nietzsche, F. (2001) *The Gay Science*. (Ed. Williams, B., trans. Nauckhoff, J.) Cambridge University Press (Original work published 1882).

Nørretranders, T. (1999) *The User Illusion – Cutting consciousness down to size*. Penguin Books.

Rábano, A. (2018) 'Aristotle's "mistake": the structure and function of the brain in the treatises on biology', *Neurosciences and History*, Vol6., No.4, pp.138-143.

Ross, D. and Ladyman, J. (2010) 'The Alleged Coupling-Constitution Fallacy and the Mature Sciences', in Menary, R. (ed.) *The Extended Mind*. MIT Press.

Ryle, G. (2000) *The Concept of Mind*, Penguin Classics.

Schopenhauer, A. (2004) *Essays and Aphorisms*. (Trans. Hollingdale, R.), Penguin Books (Original work published 1851).

Schwitzgebel, E. (2024) 'Introspection', in Zalta, E. and Nodelman, U. (eds.) *The Stanford Encyclopedia of Philosophy* (Summer 2024 Edition), available at: https://plato.stanford.edu/entries/introspection/ accessed 16/09/2024

Sellars, W. (1962) 'Philosophy and the Scientific Image of Man', in Colodny, R. (ed.) *Frontiers of Science and Philosophy*, pp.35-78. University of Pittsburgh Press.

Singer, P. (2021) 'Galen', in Zalta, E. (Ed.) *The Stanford Encyclopedia of Philosophy* (Winter 2021 Edition). Available at: https://plato.stanford.edu/entries/galen/ accessed 11/05/2024

Spener, M. (2024) *Introspection: First-Person Access in Science and Agency*. Oxford University Press.

Sullivan, S. (1995) 'Psychological Activity', *Psychological and Ethical Ideas*, pp.14-75. Brill.

Thagard, P. (2023) 'Cognitive Science', *The Stanford Encyclopedia of Philosophy* (Winter 2023 Edition), available at: https://plato.stanford.edu/entries/cognitive-science/ accessed 20/02/2025

Weinberger, J. (2000) 'William James and the Unconscious: Redressing a Century-old Misunderstanding', *Psychological Science*, Vol.11, No.6, pp.439-445.

IN DEFENCE OF THE INTERPRETATIVE SENSORY-ACCESS THEORY OF SELF-KNOWLEDGE: CONFABULATION DOES UNDERMINE INTROSPECTION FOR PROPOSITIONAL ATTITUDES

Knowing thyself, through a glass, darkly

Cases of the confabulation of propositional attitudes have been used by Carruthers (2013) as evidence that we lack introspective access to our propositional attitudes, and that instead our route to self-knowledge is via self-interpretation. In this paper I defend this position from objections raised by Andreotta (2021) that Carruthers' claimed cases of self-misattributions do not show people misattributing their propositional attitudes. In response to Andreotta's

objection for a particular case involving a split-brain patient, that the apparent misattribution involved a fabricated explanation, not a fabricated intention, I argue that this objection does not hold for cases where self-attributions are made retrospectively. Secondly, I explore the unwelcome consequences of accepting a separate objection from Andreotta that some claimed cases of misattributions are instances of propositional attitudes being influenced by perceptual cues. Accepting this objection would entail that we are mercurial beings, which plainly we are not, or that additional attitude-like mental features would be required to avoid this entailment. In defending self-interpretation accounts of self-knowledge, I seek to show that compared to direct access views they can more plausibly account for why our self-reports are swayed by perceptual cues. I go on to make the case that confabulations are a normal and everyday occurrence, and as such, our confabulations mainly go unnoticed, get glossed over, or become self-fulfilling. Consequently, the level of behavioural incongruence with self-attributions required by critics to show that confabulations have occurred is too demanding.

1 Introduction

1.1 Introspection

We attribute mental states, and in particular propositional attitudes (beliefs, desires, intentions, and so forth), to both ourselves and to others. The capacity for self-attribution of propositional attitudes is generally known as introspection, and the capacity for attributing propositional attitudes to others is known as mindreading. The standard view is that there is a principled difference between the two (e.g. Williamson 2020, p.106), in what is sometimes called the *difference thesis* (Smithies and Stoljar 2012, pp.4-6). This paper argues against the difference thesis by defending Carruthers' (2013) Interpretative Sensory-Access (ISA) theory—which holds that we only have a

single capacity (mindreading) for attribution of propositional attitudes to ourselves and to others—from objections raised by Andreotta (2021).

According to the standard view introspection provides a special kind of access to our own mental states, an access thought to be authoritative and privileged (Byrne 2005), yielding reliable outputs. This matches how our self-attributions seem to us. We seem to just know what our propositional attitudes are in an immediate way. Do I intend that I will go swimming? Yes! How do I know? I just sort of do. I did not need to work the answer out, like I would need to do if answering whether a friend intended to go swimming today, where I would draw on information like what she was carrying, what direction she was walking in, and what I knew about how much she enjoys swimming. The answer we get about our own propositional attitudes also seems particularly secure (Schwitzgebel 2024); I might be wrong when I judge that my friend intends to go swimming (perhaps I've made a mistake, or she is trying to trick me into believing that she is going swimming), but it does not seem like the kind of thing I could be wrong about when it comes to myself.

This view of self-knowledge as largely reliable and authoritative seems to match our general intuitions and experience of attributing propositional attitudes to ourselves. We seem to know things about our own propositional attitudes, with high certainty and little or no effort, that we can only find out about for others through means such as interpreting their behaviour or asking them about their state of mind. Contrast the ease, effectiveness, and certainty of considering whether you believe that it will rain

tomorrow versus trying to ascertain whether the person you spoke to last Tuesday believes that it will rain tomorrow.

The experience of self-attribution of our propositional attitudes makes it seem like we have a kind of special access, which is direct or transparent, to our attitudes. This access is sometimes considered to be a kind of perception or quasi-perception, a matter of literally 'looking within' at our own mental states (Butler 2013, pp.6-7; Lyons 1986, pp.1-3; Mandik 2014, p.148; Schwitzgebel 2024). As a result of this supposed special direct access, coming to know our own propositional attitudes is considered more secure and accurate than how we come to know the propositional attitudes of others, or how others come to know about *our* propositional attitudes (Byrne 2005; Carruthers 2013, p18-19; Kind 2016; Schwitzgebel 2024). This is held to be the case, even if our access to our own propositional attitudes can be somewhat fallible (Sellars 1956, p.415)[4].

In contrast to the standard view of self-knowledge that we have some form of direct introspective access to our propositional attitudes—wants, yearnings, suspicions, and so forth—Carruthers' (2013) Interpretative Sensory-Access (ISA) theory holds that our

[4] While philosophers generally hold that introspection is a special kind of authoritative and privileged route to self-knowledge, it is not generally thought to be infallible (e.g. Bayne and Spener 2010; Sellars 1956). The philosopher Eric Schwitzgebel, for example, reports that in some cases his wife is better than he is at determining whether he is angry about having to wash the dishes (Schwitzgebel 2008, p.629).

We shouldn't be too sceptical about introspection. Spener (2013 and 2024, p.3) thinks that some of the supposed unreliability of introspection in the literature comes from directing introspection at the wrong kind of thing, as well as confusion or ambiguity as to whether the term 'introspection' is being used to refer to our cognitive capacity to access mental states or the ways in which we employ that capacity. Just as I cannot tell how much some piece of cheese costs by listening to it or cannot work out what is wrong with a car by licking and tasting the wheels, there are some uses to which introspection cannot be usefully put. This is not a problem with introspection *per se*, rather in how we try to use it and how we theorise about it.

self-attributions are interpretative in nature. According to the ISA theory, our route to self-knowledge about our propositional attitudes is of the same kind as our route to attributing propositional attitudes to others. We use the same mental capacity to both mindread others and to do what we normally consider to be introspection. This entails that when I think about what my attitude is about going swimming, just as when working out the answer for my friend, I will draw upon a range of information to work out—interpret—the answer. Similarly, I am liable to make the same kinds of mistakes in self-attribution as I am when attributing propositional attitudes to others. A key prediction of the ISA theory is that we frequently confabulate, meaning that we will frequently misattribute propositional attitudes to themselves (Carruthers 2013, p.6). If true, this would mean that we are frequently mistaken about our own propositional attitudes.

While the ISA theory enjoys some support (e.g. Rimkevičius 2020; Scaife 2014) there are philosophers who are critical of it (e.g. Andreotta 2021; Balsvik 2017; Byrne 2012). Andreotta (2013, pp.4851-4855) makes the case that Carruthers has failed in his attempt to provide empirical evidence supporting his position that we frequently confabulate. Given the confabulation data is supposed to motivate the ISA theory, a lack of appropriate confabulation cases would substantially undermine it.

1.2 Structure of this Paper

In this paper, I seek to defend Carruthers' ISA theory from objections raised by Andreotta (2013). In Section 2, I outline Carruthers' ISA Theory and the important role of confabulation in supporting the theory.

In Section 3, I present Andreotta's objection to Carruthers' claimed confabulation data when it comes to reporting intentions and the like, focusing on the case of a split-brain patient. Andreotta argues that such cases do not show signs of misattribution of their current propositional attitudes, hence there is no evidence of confabulation. In response, I argue that we should not expect a conflict between self-reports and *subsequent* behaviour, but that we can find retrospective evidence of confabulation.

In Section 4, I give Andreotta's objection that some of the claimed cases of confabulation are actually instances of propositional attitudes being influenced by perceptual cues, and there is no reason to suppose that confabulation is occurring. In response, I argue that if we were to accept that perceptual cues were influencing our propositional attitudes rather than the reporting of them it would result in one of two unsatisfactory outcomes. Either we are mercurial beings easily swayed by extraneous factors—a view inconsistent with our apparent psychological stability—or we must posit a further set of propositional attitudes (or something like them) which are not consciously accessible but serve to anchor the content of our self-reports.

In Section 5 I give a more general response to Andreotta's objections and I discuss the kind of misattribution errors we should expect to find as part of our everyday confabulations, arguing that the evidential threshold that Andreotta is asking for in favour of systematic confabulations is too high.

Finally, in Section 6 I give a summary of the paper.

2 The Interpretive Sensory-Access Theory of Self-Knowledge

"The sorts of things I can find out about myself are the same as the sorts of things I can find out about other people, and the methods of finding them out are much the same"

– Gilbert Ryle (2000, p.149)

2.1 Mindreading and Interpretative Self-Knowledge

It is thought that via introspection and mindreading we come to know our own propositional attitudes and those of others. Propositional attitudes are mental states that consist in a stance (the attitude) towards some content (the proposition) (Schroeder 2006, p.65). Examples of propositional attitudes include hoping that it will not rain tomorrow, fearing that there might be a monster under the bed, and regretting that I tried to use some food as a doorstop. The standard view is that we have special epistemic access to our own propositional attitudes, which is often thought to be direct or transparent.

In contrast to the view of self-knowledge where we have some form of direct or transparent access to our propositional attitudes, Carruthers' (2013) ISA theory holds that our self-attributions are almost always interpretative in nature, and are no different in kind to how we attribute propositional attitudes to others. Instead of being able to 'get at', 'read', 'see', 'retrieve', or otherwise obtain some kind of 'status read-out' about our own propositional attitudes—however transparently or through a glass darkly—we work out indirectly what our propositional attitudes might be, like

constructing a story, based on secondary evidence such as our behaviour. If our propositional attitudes resided in a black box with inputs and outputs, introspective access views are about observing or getting to the contents of the box, whereas the interpretative view is about working out what could be in the box by reverse engineering what fits with the pattern of inputs and outputs that we observe.

According to the ISA theory, we do not have two distinct routes to knowledge about propositional attitudes, one for our own propositional attitudes (introspection) and one for the propositional attitudes of others (mindreading). Rather, we have a single capacity—mindreading—to attribute propositional attitudes both to ourselves and to others (Carruthers 2013, p1; Carruthers 2018).

Mindreading is the indirect and interpretative attribution of mental states to others (Marraffa 2021). We come to know about the mental states of others through a variety of means, from observing their behaviour, listening to what they say, and drawing on contextual information and our background knowledge about human psychology, social norms, and the like. Sometimes mindreading can be effortful and deliberate, but more often it is unconscious and our judgements of the mental states of others seem immediate to us (Carruthers 2013, p.12). For example, we just get a sense that someone is happy or sad about something or an impression that they intend to go out in the rain, without having to deliberately attempt to make a judgement.

How we come to mindread the mental states of others is an open question. We clearly have a folk-psychology—our everyday commonsense psychology—which enables us to think about, explain, and predict the behaviour of others in mental terms (Hutto and

Ravenscroft 2021; Luca and Gordon 2017). There are two main families of theories that seek to answer this question; the first set is based around simulating the minds of others (simulation-theory), and the other around the use of some kind of theory (the theory-theory) (Marraffa 2021). In either case, the use of folk-psychology involves theoretical posits that are tacitly or implicitly used to try and work out the mental states—including the propositional attitudes—of others (Mandik 2014 pp.141-142).

The ISA theory puts our way of finding out our own propositional attitudes on a level with how we find out about (mindread) the propositional attitudes of others. According to the ISA theory, we work out our own propositional attitudes in the same kind of way. Though in our own case we do have non-interpretative access to our sensory states and perceptual content which is globally available to us, including perceptions, mental imagery, sensorily-embedded judgements, 'context-bound' judgments (such as emotion-like states), affective feelings, working memory, and inner speech (Carruthers 2013 pp.2-4, p.49, p.69, p.94). It is from this information, along with contextual factors such as behavioural cues, that we *interpret* what our propositional attitudes are.

The difference between mindreading others and oneself is that for ourselves we have a wealth of contextual and sensory data, and we have a perspective on and history with our own propositional attitudes "since we are with ourselves all day long" (Fricke 2014, p.97). But this is not a difference in kind, such as having direct authoritative access to our propositional attitudes would be. It is a difference of the range and availability of information used in the interpretative mindreading process, albeit that difference can be substantial.

While many mental processes may be inferential (Carruthers 2009 p.123; Engelbert and Carruthers 2010, p.245), what distinguishes the interpretative route to self-knowledge is the use of "information about the subject's current circumstances, or the subject's current or recent behaviour, as well as any other information about the subject's current or recent mental life" (Carruthers 2009, p.123). It is an act of *interpretation*, of working out what a propositional attitude is likely to be based on a range of information (some of which is directly accessed, like sensory states), rather than having the answer available. So, if I am acting as though I believe that it is going to rain, and recently said this out loud, I have plentiful information to draw upon to attribute to myself the belief that it is going to rain. Conversely, if I have just met someone and wonder whether I like them or not, there may not be these kinds of robust reasons to draw upon from my sensory channels and prior history[5]. Instead, the mindreading capacity—being used for first-person attributions—could draw upon information such as my current emotions, an assessment of how our conversation has gone, my physiological states, and whether I tend to like similar people[6].

While the ISA view may seem less intuitively acceptable compared to the standard view—after all it seems as though we can know our own propositional attitudes intimately and with certainty—it is, as argued by Carruthers (2013), a better fit with the

[5] That prior history coming to the mindreading capacity via sensory channels as part of recollecting a memory.

[6] For example, in a series of experiments Dutton and Aron (1974) found evidence suggesting that under conditions of high anxiety participants were more attracted to the experiment confederate. The idea here is the cause of (non-sexual) arousal was mistakenly attributed to the confederate, rather than external conditions such as being on a shaky bridge or anticipating being shocked in a laboratory. This is a potential case of not directly accessing one's propositional attitudes and instead drawing on external and contextual factors to (mis)interpret them.

empirical data of the errors and shortcomings in our self-reports. The ISA view provides a plausible explanation for the fallibility of our self-knowledge and fits together some of the jigsaw puzzle pieces of evidence that count against direct access views, such as third-parties sometimes knowing better than we do what our propositional attitudes are (e.g. Schwitzgebel 2008).

2.1.1 Scope of Self-Interpretation

According to the ISA theory, we self-interpret our propositional attitudes based on available sensory evidence, but sensations and emotional experiences are known to us without self-interpretation. Propositional attitudes that do not surface in a sensory way or via behaviour would have little to no evidence to support the generation of an accurate self-report[7]. Self-interpretation, as considered here, is a pervasive everyday automatic process through which we make sense of ourselves.

2.2 The Confabulation Case for the Interpretive-Sensory Access Theory

According to the ISA theory we use mindreading to attribute propositional attitudes both to ourselves and to others. If this is the case, we should find (broadly) the same pattern of reliability and errors across both types of targets (Carruthers 2013, p.6). That is, the kind of mistakes we make when attributing propositional attitudes to others are the kinds of mistakes that we will make when attributing them to ourselves. That we will frequently make this kind of mistake is the "central, key, prediction" of the ISA

[7] It may be the case that we have non-interpretive access to a subset of our propositional attitudes, though this would not be in keeping with the ISA theory.

theory, that when reporting our propositional attitudes we will often be wrong in systematic ways. According to the theory, there will be frequent everyday cases where people "misattribute propositional attitude states to themselves", in what are known as 'confabulations' (Carruthers 2013, pp.6, p.157).

Confabulations are fabricated reports about people's own mental states, which when giving them the person confabulating believes to be true[8]. Confabulations are interesting and surprising because they are an instance of a person being wrong about their own mental states, where normally we suppose them to be an authority on such matters. This may in part explain why such mistakes are resistant to detection and correction.

When people confabulate, they are unaware that they are doing so. When confabulating, they make "up a story without 'realizing it'", and "fill in the gaps, guess, speculate, mistake theorizing for observing" (Dennett 1993, pp.94). We are often ignorant of the causes of our attitudes (as well as our judgements, behaviour, and so on), and in the case of confabulations we are ignorant of our ignorance yet produce a fictitious explanatory or justificatory narrative anyway (Ganapini 2019). We conjure up false explanations for our behavior that feel just as certain to us as our true explanations or any other introspective outcome (Cushman 2012, p.349).

Empirical studies lend support to the ISA theory's prediction of frequent confabulations, showing that we are often mistaken about the factors that influence

[8] The person who is misinformed or deceived is not confabulating when they repeat the misinformation or deception, nor is the person who believes and passes on a rumour, as they are not making mistakes about their own mental states.

our judgements, and (Carruthers argues) misreport our propositional attitudes. For example, cases of non-conscious influences on our attitudes and judgements are thought to be everyday cases of confabulation (Scaife 2014), like when we deny or overlook factors that influence our decisions and judgements, such as being swayed by priming, recency and positioning effects, and familiarity. Instead, we place undue weight on factors that are not influential on our attitudes and judgements (Nisbett and Wilson 1977; Schwitzgebel 2024).

A classic case of confabulation comes from a priming effects study by Nisbett and Wilson (1977), in which it was shown that when presented with identical pairs of stockings that had no distinguishing features people's choice of stockings was heavily influenced by position effects: people tended to prefer the right-most pair (Nisbett and Wilson 1977, p243). When asked, the study participants provided a range of different explanations for their choice. None of their explanations included the position effect, and when this was suggested to them "…virtually all subjects denied it, usually with a worried glance at the interviewer suggesting that they had misunderstood the question or were dealing with a madman" (Nisbett and Wilson 1977, p244). The participants were generally confident in giving their likely false reasons for their choice, such as saying what they liked about the pair they had selected, even though there was no difference between the pairs. This seems to be a clear case of misattribution.

Strikingly, in another experiment reported by Nisbett and Wilson (1977, pp.249-251), participants were asked to rate to what extent various factors influenced their judgements about another person. This was done on a seven-point scale ranging from

"increased my liking a great deal" to "decreased my liking a great deal". A second group of participants was asked to rate how much the first group of participants would be influenced by the same set of factors. The ratings from the two sets of participants were found to be highly correlated, in that both groups gave similar ratings for how much the factors would influence the first group. This was despite most of the first group's ratings being inaccurate as to how much the relevant factor did influence their judgements (Nisbett and Wilson 1977, p.250). This is a case of the same kind of mistake being made from a first-person perspective as a third-person perspective, about the same thing. It suggests our insight into our own judgements is not that different from the insights we have into the judgements of other people. This is what we should expect to see if we use a single capacity for self and third-party attribution of propositional attitudes (Carruthers 2013, pp.339-345). If this is the case, it raises a challenge for direct access views to deal with as to why a direct route to self-knowledge leads to the same kinds of errors as third-party interpretation. Such a challenge is all the more difficult because direct access views hold that our route to self-knowledge is particularly authoritative, privileged, and secure.

The ISA theory does not claim that we misattribute propositional attitudes to ourselves all the time, but that there should be a pattern of errors that align with the mistakes we make in third-party mindreading. They won't be quite the same mistakes, as we have much more data about ourselves to inform (or mislead…) our self-attributions, but we should see the same kinds of mistakes with the same underlying causes, because the same capacity and broadly similar information is used in both cases. Whereas introspective access accounts will not properly account for the pattern of

these everyday confabulations[9], under the ISA theory we should expect confabulations to occur whenever people are "presented with the right sorts of misleading behavioural and/or sensory data" (Carruthers 2013, p.6).

It is worth keeping in mind that for some of our self-attributions we will draw on a consistent and strong body of evidence, which makes misattributions or variations in our self-reports relatively unlikely compared to misattributions in third-party mindreading. For example, on the matter of whether or not I believe that a particular prominent politician is a force for good, it seems likely that subsequent self-reports on the matter are more of a recollection of what I previously thought or reported, than an interpretation. I can recall my previously stated views, my declared reasons for giving them, and so on. This is compatible with the ISA theory but does highlight that in searching for instances of confabulation we need to focus on self-attributions which are relatively fresh, novel, or perhaps less weighty and less memorable than matters like whether we believe that a particular politician is a force for good or not.

As frequent and systematic confabulations are the key prediction of the ISA theory, anything that challenges the presence of confabulations or how they fit with the theory is a substantial threat to it. Thus, any proponent of the ISA theory needs to take Andreotta's (2021) objections seriously. Andreotta argues that there are no "noteworthy patterns" in the data presented by Carruthers (Andreotta 2021, pp.4854), and that our route to self-knowledge is non-interpretive and non-confabulatory. Based on this, Carruthers' claimed cases of confabulation do not give us reason to doubt that

[9] At least not without proposing dual processing routes whereby some attempts at self-knowledge involve direct access and others do not (Carruthers, 2013 pp.333-339).

we have access to our propositional attitudes. While such a move, if successful, would not refute the ISA theory, it would substantially undermine the case for it.

3 Confabulation of Intentions

In this section I set out Andreotta's (2021) objection to Carruthers' claimed confabulation data where those cases relate to reporting intentions, focusing on the case of a split-brain patient. I outline the interpretive view, give Andreotta's objection, and then respond to that objection.

3.1 The Split-Brain Patient

This example of confabulation relates to propositional attitudes being reported to explain behaviour. Studies have been carried out with split-brain patients, whose corpus callosum—which connects the two brain hemispheres—has been severed, which substantially inhibits information transfer between the two. In these studies, information is presented to only one of the split-brain patient's eyes, and therefore only to one of their brain hemispheres[10]. In one such study (see Andreotta 2021 and Carruthers 2013 pp.39-40), an instruction "Walk" was flashed to the right hemisphere (via the left eye) of a split-brain patient. While the right hemisphere has some language comprehension, it is the left side that carries out language production. The participant got up and started walking out of the testing van he was in. When asked where he was walking, he gave the good-faith self-report of his intention as "I'm going to get a Coke from the house". The story here is that the brain hemisphere responsible for giving the explanation did not have access to the "Walk" instruction, so was unable to give the true answer. Rather than respond along the lines of "I do not know", the reporting

[10] There may well be some information transfer between hemispheres via other routes, but generally findings indicate that there is much reduced information transfer, as would be expected from a brain with a severed corpus callosum.

hemisphere in good faith interpreted from the situation and reported a plausible reason for getting up and walking towards the house. Instead of saying "I do not know", so the story goes, the participant made use of the information that was available to them in generating the self-report, something along the lines that they were thirsty, they quite like Coke, and Coke was available to them nearby (and getting up and walking to fulfill the intention to get a Coke is a plausible explanation that fits the situation…), so they were off to get a Coke. These bits of information are unrelated to what caused them to get up, but from them a plausible response has been created, one that coheres with the available evidence. This example is instructive, because due to the individual's split brain we can be confident that there has been a fabrication involved directly or indirectly in the self-report (due to a lack of access to the real reason for getting up and walking: the instruction to walk).

Both Carruthers (2013, p.95) and Andreotta (2021, p.4865) accept that the patient is actually intending to get a Coke at this point of the scenario, and both accept the intention has come about as a confabulated rationalisation (Carruthers 2013, p.40 and Andreotta 2021 p.4865). Where they disagree is whether the self-report the patient gives—that he intends to get a Coke—is a confabulation or not. Carruthers thinks that it is a confabulation, Andreotta argues that it is not.

3.2 The Interpretative View

Going to get a Coke was plainly not really the motivating intention for the participant to get up and walk—they did so in response to the "Walk" instruction—and so (as we are assuming they're being honest) they have confabulated. While an example involving a

split-brain patient is quite exotic, it gives us a clear 'zoomed in' picture of a phenomenon which also takes place as an everyday occurrence.

Carruthers (2013 p.40) tells us that the experience of split-brain patients reporting their propositional attitudes is the same as for people without split brains. Their reports are "delivered with all of the confidence and seeming introspective obviousness as normal" (Carruthers 2013, p.40). In the case of getting the Coke, the participant provided his answer "smoothly and unhesitatingly" (Carruthers 2013, p.42) just as we would expect from someone with a non-split brain. Despite being made aware of the impact their split-brain has on their ability to provide accurate self-reports, the split-brain experiment participants continue to confidently provide confabulated self-reports without any seeming reference or consideration of their condition (Gazzinga 1995 via Carruthers 2013, p.40). If the experience of confabulating versus giving true self-reports is the same, it lends further support to the idea that we (sometimes) confabulate on an everyday basis and do not notice.

Carruthers thinks the report of the intention to get a Coke is confabulated, because the reported intention is not the reason the person got up, and it was not there until the reason was confabulated, so it is part and parcel of the confabulation event. In Carruthers' view the intention to get a Coke follows from the confabulated reason; it becomes self-fulfilling (Carruthers 2013, pp.94-95). Once we have declared why we're doing something, then we are (at least in part) committed to it and likely to behave accordingly (Carruthers 2013, pp.95). Additionally, the declared reason given for getting up—to get a Coke—will likely be based on an interpretation of relevant factors

like feeling a little thirsty, liking Coke, and having Coke available, and so it should come as no surprise why people would act in conformance with their confabulations.

The picture then, under the interpretive-access view, is that the subject self-interprets themself as having a particular intention and then reports it. This activity would also make the interpreted propositional attitude more generally available for future self-reports, as information about the purported intention and subsequent related future behaviours are sensorily accessible (e.g. the participant *said* he wanted to get a Coke)[11]. We should expect the subject to continue to report the same intention and aim to act upon it, and for there to be little in the way of behavioural mismatch. But they are still interpreting themselves to have the intention, rather than accessing their propositional attitudes. Thus, we would expect self-reports on this matter to be largely consistent over time, even if some sensory and contextual information changes. This is Carruthers' view: that when we seek to self-report what occurs is **interpretation-then-report**, and in this case the report was a confabulation. Until the individual was questioned there was no intention to get a Coke, the intention was fabricated to give an answer to the experimenter about why they got up. The intention comes into existence as part of attempting to report a reason for why they got up.

3.3 Confabulated Intention, Accurately Reported?

In contrast to Carruthers, Andreotta sees the causal reason for the propositional attitude and occurrent intention as distinct. Andreotta (2021, pp.4864-4865) agrees

[11] This kind of story is an example of how once content is confabulated, it becomes self-reinforcing. 'Of course I want a Coke, I just said so", etc.

with Carruthers that the participant confabulated when he reports that "I am going to get a Coke from the house", after all we know that is not the reason why the participant started walking. But Andreotta (2021, pp.4865) thinks only the causal explanation for the participant's behaviour is a confabulation, that their reported intention is a correctly reported propositional attitude, or at least there is no reason to doubt that is the case.

Andreotta (2021, pp.4865-4866) thinks that the case does not provide an example of a confabulated self-report of a propositional attitude, because we have no reason to doubt that the participant really does have an intention to get a Coke at the point at which he reports his intention. His behaviour is congruent with wanting to get a Coke. As a result, there is no error in his self-report, or at least no reason to think that there is an error, and so no evidence of the self-report being a confabulation (though there is evidence for his confabulating the reason why he got up to walk out of the testing van). If we attempt to report on our intentions, and what we report ('intend to get a Coke') matches our behaviour (directed at getting a Coke) then that *is* our intention, or at least the self-report is not a confabulation, even if its content originally came about via a confabulation. Andreotta accepts the causal reason for getting up was confabulated but maintains that nevertheless the occurrent future-orientated intention being reported is correct and the self-ascription of such an intention is therefore not a confabulation.

Andreotta's (2021, p.4865) view of this case is that the causal explanation the participant gave ("I'm going to get a Coke from the house") when asked where he was

going was fabricated into existence as part of attempting a self-report, and this leads to one or more relevant propositional attitudes which can be subsequently introspectively accessed. The reason for getting up was confabulated which results in there being an intention which is accessed and self-attributed. Under this view we fabricate a propositional attitude, then access it, and then report it; **generation-access-then-report**, rather than interpretation-then-report[12]. Under Andreotta's view there is no reason to consider the reported intention a misattribution, because at that point the participant really did intend to get a Coke and would have acted upon that intention if he had not been stopped.

In summary, both Carruthers and Andreotta believe that when asked to give a reason for getting up and walking about, the subject confabulated a reason (an intention: they wanted to get a Coke). For Carruthers, the intention was a confabulation brought about by an interpretative process, and then that interpretation was reported to the experimenter. The confabulation in this case is the self-attributed intention. In contrast, Andreotta believes the causal origin of the intention is a confabulation, and that the resulting intention is then directly accessed and self-attributed.

3.4 Explaining Intentions with Retrospective Confabulations

Andreotta has argued that while the causal origin of an intention may have come about from a confabulation, we have no reason to doubt that intention is subsequently

[12] The generation step here could be 'interpretation', even under direct-access views. However, I am not committing Andreotta to the idea that the confabulated intention to get a Coke comes about by interpretation, instead I'm trying to emphasise Andreotta's view seems to imply an additional step not present in Carruthers' view, fitting 'access' between generation/interpretation and reporting.

accessed as there is no incongruity between people's self-reports and behaviours in the cases raised by Carruthers. Therefore, the cases do not provide evidence of a pattern of confabulation in support of the interpretative view of self-knowledge (Andreotta 2021, pp.4865-4866). My response to this is to argue that there is an error, but that the error is retrospective, with the incongruity being between people's self-reports and what those self-reports are meant to be about.

In the case of the split-brain patient getting up and reporting that he is getting a Coke, the idea is that while an intention to get a Coke was not the reason why he got up, when asked he gives his actual current intention: to get a Coke. Between getting up and giving the report, the patient has acquired the intention to get a Coke, and both Carruthers (2013 p.95) and Andreotta (2021 p.4865) are of the belief that he really would have carried on getting a Coke if he was not interrupted by the experimenter. Andreotta considers this to be a case of a successfully accessed propositional attitude (or, at least, that there is no evidence that this is not the case) as the patient's behaviour is in line with this self-report. This lack of error indicates to Andreotta that there is no evidence of the reported intention being a misattribution, so there is no reason to suppose a confabulation, and so this case and those like it do not support the ISA theory. However, we can say that there *is* error here. While Andreotta makes a useful distinction between the causal reason for a particular propositional attitude and the propositional attitude itself, we can be confident in the Coke case that the intention to get one was originally a misattribution. At time *t*, the individual did not intend to get a Coke, he got up because he was instructed to. At time *t+1*, he was asked why he had gotten up and started walking. At time *t+2*, he answers "I'm going to get a Coke from

the house". This self-attribution is false and seems confabulated, at least with respect to time *t*. Even if we accept that at time *t+2* that his answer is a true self-report of his future-directed intention, it is not plausible to hold that it is correct with respect to time *t*. Time *t+1* is the triggering event—being asked what he was doing—that led to the confabulated content and self-report. Before the triggering event *(t+1),* he did not have the intention to get a Coke. After it, he did. The self-report at *t+2* is not a misattribution for him at *t+2*, but it is a misattribution with respect to *t* and *t+1*.

In his argument Andreotta is focusing on *t+2* and saying with respect to *t+2* the self-report is correct. Further, that there is not any evidence, and won't be any evidence, that the declared intention is not the participant's actual intention, because their behaviour will align with the self-report. The patient will go and get a Coke, fulfilling their declared intention. But this is perhaps too fast a move. The interpretive-access theorist can point to the self-fulfilling aspect of giving a self-report. We believe them to be true, and we want to act consistently, particularly in the presence of others so that they consider us to be reliable and trustworthy (Murphy-Hollies 2022, p.114 and Ganapini 2019, pp.196-197). Under both interpretive and direct access views we should expect the participant to appear to act rationally and aligned with his self-reports, so we should expect him to go and get a Coke, or, if he does not, to give us a good reason as to why he has changed his mind.

For the argument to work then, we must split out the confabulated causal reason for getting up and walking and the new occurrent intention. The direct access theorist then needs to make the case for why we should judge whether the self-report is

accurate based on a future-orientated perspective. In doing so there needs to be a good reason why the confabulated intention given as a retrospective answer should not be considered. This may not be so straightforward, and in many cases it will be the retrospective answer that is relevant, making it clearer that a confabulation has occurred.

When we ask someone about what they are doing and why, often we are asking retrospectively. We are asking "why are you doing this thing that I have observed you doing?". We are not asking—particularly after observing someone doing something for which we want an explanation for—"what are you aiming to do in a few minutes time?". It would be rare for us to mean "in response to my question I want you to come up with an intention that you did not previously hold but matches your past behaviour and will likely match your future behaviour". This is true of the 'getting a Coke' case. When the split-brain patient was asked why he was walking out of the testing van, the question in play was "what propositional attitudes are you acting upon that have led you to get up and walk?", and we want an answer that addresses that question. Or perhaps we are asking "what are you doing right now and why?". In both cases, we are after an answer that is true and satisfies our curiosity about his behaviour prior to the point of asking about it, or at the point in time of asking. That is not what we get. The reason given, and the related propositional attitudes, are not well-grounded and do not capture what brought the behaviour about. They are a misattribution, and this shows that the participant was unable to successfully access the relevant propositional attitudes and was not aware of his failure to do so.

I'm not claiming that *every* question about people's intentions and similar propositional attitudes are meant to be answered retrospectively. In this case though, the split-brain patient was sitting down inside a 'testing van' taking part in a scientific study as a participant. He got up and started to leave the van. This is presumably an unusual departure from normal behaviour for a participant inside a testing van partway through a scientific study. When asked why he had gotten up and started walking, the question was about his departure from his previous behaviour, i.e. what made him get up, or about what he was doing at that point in time. In this specific case, it is highly likely that the question was asked retrospectively because the experimenters wanted to prompt a confabulatory response, knowing that the participant would be unable to report the real reason why he got up and started walking. But even if this were not the case, the question would clearly be about the participant's departure from his previous behaviour, i.e. what instigated the change.

We know that the true answer as to why the participant got up was not because he wanted to get a Coke. It can be argued that at the point of answering (*t+2*) the patient really does want to get a Coke, so he gave a true report of one of his propositional attitudes. But nevertheless, it should be considered a misattribution. The participant did not have an intention to get a Coke until *after* he was asked why he had got up and started walking, and so even if the propositional attitude has been correctly accessed and reported, it is a misattribution because it is not relevant or correct to report it in response to the question. It is a false, fabricated response about the relevant mental states given in response to the question. While this position does not directly support

an interpretive access view, it does undermine the case that we have direct and special access to our propositional attitudes.

A defender of direct access views critical of this response that the self-report is a confabulation because it is retrospectively false could object and suggest that whether or not someone 'changed their mind' at the point of being asked to give a propositional attitude is beside the point. The critic can emphasise that the participant is reporting their propositional attitude correctly by virtue of the relationship they have with their own propositional attitudes, and there are no signs to suggest it is a misattribution for them to declare an intention that they're going to get a Coke.

However, even if this were the case, there is still scope for the self-reported propositional attitude to be wrong with regards to what has happened in the past. If a self-reported propositional attitude conflicts with the recent past behaviour (and absent any compelling justification for why their behaviour did not match), then there is space to say that an individual is wrong about their propositional attitudes. We would have further grounds to say that the individual is wrong if their claims just do not make sense or some other propositional attitude would have much more explanatory power. Thus, if someone were acting melancholy and morose, we would not think they were correct when they say that they are happy. Or if a colleague gets up from their desk after the fire alarm sounds and gathers at the muster point, if asked why they are at the muster point they answer "I wanted to get some exercise by going for a walk", we can doubt that this self-attribution is correct. Hence, in the case of the split-brain patient declaring their intent to get a Coke, we have good reason to believe

that a misattribution has occurred even if they are correct about what their occurrent propositional attitude is at the point in time that they self-report. The content of that report has been fabricated in response to the question, and so we should not see the situation as being of an unambiguous case of an accurate self-report as Andreotta seems to suggest.

Lotte the Philosophy Student

To give another example, imagine someone called Lotte who wants to study philosophy. She is motivated to do so by her love of wisdom. She believes studying philosophy will make her wiser, which she values as an end in itself. When Lotte talks to her career advisor and is asked "Why have you chosen to study philosophy? Is it because it will boost your career prospects?", the true answer will be along the lines of "because I want to be more wise, I believe that studying philosophy will make me more wise; therefore, I want to study philosophy". If Lotte instead answers "because I believe that it would boost my career prospects" she would be giving a false report. She did not have any propositional attitudes about philosophy and career prospects prior to the question, so they're not relevant as an answer to the question, and they certainly had no relevance to her behaviour up until the time she spoke with her career advisor. It may well be the case that Lotte had a sudden conversion and now wants to study philosophy because she has come to believe that it will boost her career prospects. If so, giving an answer about her career prospects would be appropriate if she was asked about her future, but she's not being asked about the future, she's being asked a

retrospective question about why she made a choice to study philosophy prior to talking to the advisor.

Now, in this case Lotte's propositional attitudes have not come about via confabulation, but we can see that a mistake is being made if Lotte's reply is about her new belief about her career prospects. Even if all her subsequent behaviour (and even past behaviour) fits with being motivated by wanting to boost her career prospects, giving an answer about boosting her career prospects in response to a question about why she had chosen to study philosophy is a misattribution. If Lotte gives a retrospective answer about wanting to be wise she is giving a rationale, and it has (correct) explanatory power about her actions to date. Giving an answer about her beliefs about career prospects is a false self-attribution for explaining why she has been doing what she has done, as before the advisor raised the idea it was not part of her propositional attitudes relating to studying philosophy. Talking of being motivated by career prospects only makes sense as a prospective answer. Similarly, in the Coke case, if a direct access theorist is looking for incongruities between self-reports and subsequent behaviour, they are looking in the wrong temporal direction for evidence of confabulations.

Whether we should understand questions about someone's propositional attitudes as prospective or retrospective depends on the question and the context. In the case of Lotte, the career advisor is clearly asking what has led Lotte to consider a career in philosophy, even though at that point discussing her career prospects would be germane. Similarly, in the case of the split-brain patient the question is about what

motivated them to get up and start walking out of the test van: it is a retrospective question.

A direct access theorist could still insist that we should not consider the Coke-seeking split-brain patient's self-report to be a case of confabulation. They could argue that I am pointing to error with respect to mental states in the past, and that at the point of self-reporting it really was the case that the individual intended to get a Coke (and Lotte really does want to study philosophy to boost her career prospects), therefore there is no evidence of a misattribution with respect to the relevant occurrent intentions.

An issue with such an objection is that the view implies a process whereby propositional attitudes are confabulated and then there is some separate process which 'retrieves', 'reads', or 'takes a look' at the propositional attitude to report on it. That is, **generation-access-then-report**. This is an expectation linked to direct access views, whereby there is a *thing to access*, and therefore a step that exists between the creation of the thing and reporting the thing. It requires a strict division between past causes and occurrent intentions, which needs defending. But this view claims that even if what is subsequently accessed was confabulated, it does not count against introspective or direct access views as long as what is accessed and reported is congruent with subsequent behaviour. But that congruence will likely arise because of the self-fulfilling nature of declaring plausible propositional attitudes which fit the available evidence. On this basis, there will rarely be prospective evidence of

incongruence between self-report and behaviour even when there definitely has been a confabulation.

The interpretive-access view seems to make more sense of what is going on in cases where we are asked to report our intentions, particularly when the question is directed retrospectively. Under the interpretation view the generation or interpretation of propositional attitudes for a causal explanation and accessing occurrent propositional attitudes are not necessarily distinct steps. Rather, the interpretation view of the process is that when self-reporting our relevant propositional attitudes are interpreted; **interpretation-then-report**. That's where the error can occur, and that's where it seems to have occurred in the case of the split-brain patient reporting he intends to get a Coke. He did not intend to get a Coke before being asked, and the question was not about what he wants to do in the future, it was about his current state and recent past, hence his self-report was a misattribution.

4 The Influence of Perceptual Cues on Self-Reports

In this section I consider an objection from Andreotta that some of Carruthers' claimed cases of confabulation are instances of perceptual clues influencing our propositional attitudes. It focuses on an experiment involving shaky handwriting, which involves external stimuli leading to a direct or indirect change in a self-attribution of propositional attitudes. Carruthers claims it as a case of misattribution, but Andreotta argues that it is a case of contextual cues influencing propositional attitudes, which are subsequently accessed successfully. I argue that if we were to accept Andreotta's position, that if our propositional attitudes can be influenced as readily as the position entails, then we would be far more variable, more mercurial, than we actually find ourselves to be.

4.1 The Shaky Handwriting Case

Carruthers (2013, p.344) discusses an experiment in which participants were asked to "think about and then write down three good or bad qualities that they thought they had as potential professionals" (Briñol and Petty 2003, p.1133). The participants came up with qualities like "polite", shy", and "rigid" (Briñol and Petty 2003, p.1133). The participants were randomly allocated to groups. Half of the participants were asked to write their qualities down with their left hand, the other half with their right hand[13]. The immediate outcome from this experimental manipulation is that those who wrote

[13] The experiment was a 2x2 factorial between-participants design. The other condition was whether the participants were asked to write positive or negative qualities. The discussion here focuses on just the observed differences between those who wrote with their right hand and those with their left hand.

with their left hand—their non-dominant hand[14]—produced lower quality shaky handwriting. All the participants were then asked to rate their confidence in their stated qualities on a scale of 1 (not at all) to 7 (extremely) (Briñol and Petty 2003, p.1133). Those participants who produced the off-hand shakier writing systematically expressed lower confidence in their own professional qualities than those who wrote with their dominant hand.

This experiment gives us a relatively clear case of reporting propositional attitudes. The participants provided a set of qualities and then were asked to indicate their degree of confidence that they had those qualities. So, we have a propositional attitude of the form 'I am <level of confidence> that I possess <quality X>'. In the experiment there were no causal explanations being generated, no attempts to explain behaviour, just an indication of the participants' confidence that they possessed a particular quality on a seven-point scale. In this case, they were only indicating confidence on a scale, but to be able to do so they would need to report on relevant propositional attitudes (e.g. asking themselves "How confident am I that I am polite…?").

The result of the experiment is that those who wrote with their non-dominant hand and produced lower quality handwriting also expressed lower confidence in their professional qualities than their peers. A systematic difference in their self-reports influenced by nothing more than writing with different hands and looking at poor or normal quality handwriting. This is a clear example of an extraneous (i.e. seemingly unconnected and irrelevant) sensory input contributing to our metacognitive

[14] Only one of the participants was left-handed, and their data was excluded from the analysis.

judgements, leading to a difference in the magnitude of an attitude between groups (e.g. going from “I am rather confident <that I am polite>” to “I am only slightly confident <that I am polite>”).

Similar effects of seemingly extraneous factors influencing reported judgements are observed in other experiments. In these experiments participants carry out supportive or detractive behaviours relating to their judgements, which results in a shift in their reported degree of belief, confidence, or endorsement of those judgements. For example, the amount which people are paid to write an essay defending a proposition will influence how much they report believing that proposition (Linder, Cooper, and Jones 1967). Similar effects can be achieved if participants shake or nod their heads, behaviorally confirming or rejecting their reaction to a message (Wells and Petty 1980).

4.2 The Interpretive View

To Carruthers, the shaky handwriting experiment provides an example of participants reporting a concurrent rather than a past judgement (Carruthers 2013 p.344), and involves reporting a propositional attitude (their confidence in their professional qualities). In contrast to the discussed split-brain patient case, this experiment does not involve confabulating some causal explanation. Given that in the experiment the influence of sensory based information (the handwriting quality) leads to a systematic difference in reported propositional attitudes, then for Carruthers it supports an interpretive view of self-knowledge, i.e. self-reports are not based on accessing a propositional attitude, because external factors are affecting the outcome. For the shaky handwriting case, the interpretative view provides an explanation of how the

sensory information about the shakiness of the handwriting (which is accessed) influences the self-reports; that it, along with other contextual factors, is taken into account when interpretating what the relevant propositional attitude is.

The ISA theory is further reinforced by evidence from a later stage of the experiment which used independent judges. The judges looked at the participants' writing and rated how confident they thought the participants were that they possessed the qualities they had written down. The outcome was that the judges also considered that those who produced the shaky handwriting had lower confidence in their stated qualities than those in the regular handwriting group. This shows the shaky handwriting having the same effect on attribution of propositional attitudes across self-reports and third-party mindreading. This looks like self-reporting and third-party reporting (mindreading) processes responding in the same kind of way to the same sensory stimuli, which suggests a commonality between them. This is what we would expect to see according to the ISA theory.

4.3 Our Judgements are Influenced by Seemingly Unrelated Factors

Andreotta (2021), contra to Carruthers (2013 pp.344), thinks the shaky handwriting experiment does not support an interpretative account of self-knowledge. He suggests the participants could be correctly reporting their judgements, but those judgements are unconsciously influenced by the quality of their handwriting. He gives two supporting reasons for this position (Andreotta 2021, pp.4858-4859). The first reason Andreotta gives is that empirical evidence shows that our judgements are influenced

by a variety of seemingly unrelated factors, which gives us reason to believe the experimental effect caused by the shaky handwriting need not have come about by confabulation. The second is that the participants do not behave as if their confidence declarations are false (i.e. there is not a conflict between reported propositional attitude and behaviour), and so we have no reason to consider them a misattribution. I will address the first reason in what follows, and the second reason in Section 5.

Andreotta argues that we should not take the shaky handwriting experiment to give evidence in favour of confabulated self-reports because our judgements are influenced by seemingly unrelated factors. This does seem to be the case. The shaky handwriting experiment and numerous other experiments provide ample evidence to suggest that our judgements are often swayed by a wide range of extraneous stimuli. Andreotta (2021) suggests that in the shaky handwriting experiment there may be an unconscious bias in play, associating neat handwriting with truth and not taking the content of messily written sentences seriously. He also suggests that the participants may "not have strongly formed opinions about their professional attributes, [so] their judgements may be easily influenced by how messy the handwriting is" (Andreotta 2021 pp.4858-4859). Note that this suggests that Andreotta considers it likely the participants already had relevant propositional attitudes, and that the experimental manipulation changed them. Clearly extraneous external factors can influence our judgements, and Andreotta (2021, p.4859) suggests this does not give us any reason to suppose we cannot access our propositional attitudes.

Notice that in the case of the shaky handwriting experiment, the threshold for extraneous factors to influence our confidence in our own abilities seems rather low. A case of mere bad handwriting negatively impacted people's reported confidence in their own abilities. This fits with the point that Andreotta (2021, p.4866) makes when he says that our occurrent propositional attitudes are "fragile" and "arise suddenly and mysteriously". Accordingly, I will call this view, which has reliably accessed propositional attitudes subject to change by trivial extraneous factors, the 'fragile attitudes' view.

4.4 Against The 'Fragile Attitudes' View

While Andreotta is not making a positive case in favour of the fragile attitudes view, he ends up defending some form of it via his objections to the ISA theory. However, problems arise if we are to accept that the observed phenomena in a range of claimed confabulation cases, including the shaky handwriting experiment, are due to propositional attitudes being influenced by such unconnected perceptual cues, rather than the self-reports being influenced. If we were to accept this position, it would result in us being mercurial beings who would be difficult for others to understand via mindreading.

My response to Andreotta's objection, which I expand upon below, is to explore the implication of accepting this 'fragile attitudes' view and argue that it is incompatible with the reasonably good track record of our mindreading ability and the relative psychological stability we find ourselves and others to have. Andreotta's account has our propositional attitudes being successfully accessed, but after they have been

influenced by perceptual cues. This influence includes trivial in-the-moment extraneous factors such as shaky handwriting changing our propositional attitudes. The result of this would be that our propositional attitudes would be so highly variable that we would be rather mercurial beings. Far more mercurial than the degree of relative predictability that successful mindreading requires. Yet, our mindreading ability is rather successful; at times we can be as good or better at accurately attributing propositional attitudes to others as we are to ourselves (e.g. Briñol and Petty 2003, Schwitzgebel 2008 p.629). Therefore, as we are good targets for third-party mindreading, the 'fragile attitudes' view is unlikely to be correct.

Direct access views need to account for how it is that the content of our self-reports is influenced by perceptual cues, yet we remain psychologically stable over time. If the perceptual cues are 'changing our minds' in the now, there needs to be an account for how we end up with the longer-term psychological stability and consistency that we enjoy. They also need to account for how perceptual cues can influence our self-attributions and third-party attributions in the same way, which is not what we should expect if they are carried out by distinct capacities where one involves interpretation and the other does not.

Fragile Attitudes

Under Andreotta's view, we have reliable and direct access to our propositional attitudes, which may have been confabulated or influenced by perceptual cues. In effect, it is our propositional attitudes which can be unreliable, rather than the means of accessing them. While we should expect (and want) our propositional attitudes to be

responsive to external factors (e.g. so that we can adapt and learn), under Andreotta's 'fragile attitudes' view our propositional attitudes end up being *too* responsive to stimuli, influenced by trivial extraneous factors like the quality of our handwriting. Given that under the fragile attitudes view it is our propositional attitudes which change, the implication is that this change will not be some in-the-moment influence but will last until some other intervention or stimuli influences our relevant propositional attitudes.

If our attitudes were this fragile, due to being too responsive to stimuli, we would be quite changeable and easily swayed from the target of our desires and intentions. If mere shaky handwriting can lead to lower confidence in our professional abilities, then we might expect such outcomes as each new delightful dish we encounter to be our preferred dish. Or, if for us to act more warmly towards people when we hold a warm cup of tea.

While we should accept that we are sensitive to external factors, we are also relatively stable and predictable, which does not match the characterisation of our propositional attitudes as arising "suddenly and mysteriously" or of them being fragile (Andreotta 2021, p.4866). We can admit there is mystery in terms of how little we know about the mechanisms involved, but not that they are arising mysteriously when considered in terms of folk-psychology. We have coherent and predictive ways of talking about people's propositional attitudes which play out. 'She does not like to get wet, and as she believes that it will rain today, she will want to take her umbrella with her', is a recognisable and coherent set of predictive statements about the mental states of

another. This is not a good fit with the idea that our propositional attitudes are fragile, except in relatively trivial and small ways.

Trivial Extraneous Factors Leading to Propositional Attitude Change

The shaky handwriting experiment demonstrates the influence of seemingly trivial extraneous stimuli leading to a systematic difference in reported propositional attitudes. According to Andreotta's view, those seemingly extraneous sensory inputs—such as shaky handwriting—are not merely influencing our self-reports, they are changing what we are reporting on. So, when I attempt to give an answer to what my preferred flavour of ice cream is, the attempt to check my preference may change that preference based on extraneous factors in play at the time. Perhaps presenting me with pistachio ice cream (yuck) would never be enough to shift my stated preference away from caramel (yum), but maybe if I thought enough about *toffee* flavoured ice cream, or the right kind of music was playing[15], then I would start to doubt or shift my preference away from caramel. Or, in the case of shaky handwriting, perhaps if we often wrote with our off-hand and looked at the results when delivering self-reports related to what we had written, we would have a general crash in our confidence compared to other people[16]. While we might get used to such influences and adjust, we would still be vulnerable to novel experiences and influencing factors.

If our self-reports can be swayed to such an extent by such trivial factors, it creates a problem for direct access views. Under such a view, extraneous factors (e.g. the quality

[15] To pick an extraneous factor that might influence such things.

[16] Or, presumably, those like me who have terrible handwriting would be lacking in confidence with regards to just about everything we write… this does not seem to be true.

of our handwriting) lead not only to a change in what we report about our propositional attitudes, but they also change our actual propositional attitudes. At first pass, this does not seem problematic. It accounts for how our propositional attitudes seem somewhat changeable due to extraneous factors, such as priming effects or stereotype effects (see, e.g., Doyle and Voyer 2016). Plus, we should want our propositional attitudes to be revisable due to contextual factors. We do not want to be rigid and unchanging in our preferences! We *do* want to be swayed from our intentions if they prove difficult to meet. We *do* want to change our beliefs in the face of counterevidence. We *do* want to eventually overcome our dread of the monster under the bed. We want to update our propositional attitudes as we learn and adapt to our experiences and new knowledge. But the impact of the fragile attitudes view is far more broad ranging; it would not be limited to cases like revising our preferred flavour of ice cream after sampling new flavours, including seemingly trivial and seemingly irrelevant stimuli like seeing our own shaky handwriting resulting in having lower confidence compared to other people. Our attitudes would be changing all the time, swayed by external factors, like we were blowing with the wind.

Mercurial Beings

Recall that the fragile attitudes view is a view about propositional attitudes in general, the idea that in the shaky handwriting case the participants' self-reports were influenced by perceptual cues is not given as an isolated case. According to Andreotta (2021, p.4866) our occurrent propositional attitudes are fragile, and our beliefs, desires, and intentions can "arise suddenly and mysteriously".

If the 'fragile attitudes' view was true, the result would be that our propositional attitudes would be rather mercurial. Without our propositional attitudes being relatively fixed, we would change quite substantially as we are swayed by a barrage of extraneous trivial factors in the environment, leading us to acquire and lose different propositional attitudes without much internal coherence. If merely writing a statement in shaky handwriting leads to a negative shift in our attitudes, repeated such instances (or similar activities) could potentially substantially change our propositional attitudes rather quickly. We might be quite a different person from one day to the next, with little psychological continuity. Whereas under an interpretive view, it would only be our self-reports that are directly influenced, and so our propositional attitudes can remain largely stable. This would also mean that as our self-reports get influenced by different stimuli they would only be influenced away from a stable starting point, whereas if our propositional attitudes were being directly changed, there would not be this consistent 'centre of gravity' and so our propositional attitudes could drift substantially due to repeated influences.

The proponent of the fragile attitudes view could say that perhaps there is some kind of baseline set of propositional attitude states that serve as a reference point. This baseline would enable our propositional attitudes to be relatively stable over time despite being swayed by trivial extraneous factors. This would allow for some of our propositional attitudes to change and thus for us to be subject to mercurial change, but for that change to be short-lived because there are enough other propositional attitudes that remain stable that can serve as a baseline reference. If this were so, it would fit with the general stability of our mental states which we observe. It might be

the case that access attempts do lead to a change in propositional attitudes, but one's attitudes overall do not vary much over time beyond a short-term effect at that point in time. This story fits better with our experiences and the success of mindreading. Under this view, our propositional attitudes would never really go too far astray from the baseline set due to trivial extraneous sensory information because they tend to revert to some norm. They are in effect anchored, restrained from straying too far or for too long from a central point due to extraneous factors, but changeable due to more rational and deliberate reasons. Such an arrangement would save direct access views in a way that accounts for why our self-reported propositional attitudes are quite changeable due to extraneous factors, while also providing some stability in our propositional attitudes in a way that reflects our general experience. Such anchoring could be achieved by features such as unconscious biases, which are somewhat like inaccessible propositional attitudes[17], or unconscious processes that lead to some degree of coherence amongst our various propositional attitudes, reigning in those that become outliers.

It may be the case that some core set of propositional attitudes are more fixed and resistant to change, but if non-trivial propositional attitudes like our confidence in our potential professional abilities (not to mention the impacts of stereotype threat and the like) are subject to change from trivial factors, then we would be rather changeable creatures. The proponent of the fragile attitudes view would likely need to make a case that it is only a subset of our propositional attitudes are fragile, perhaps those that are

[17] Unconscious biases, in their propositional attitude-like role which they can play while being inaccessible to self-report mechanisms, pose their own potential difficulties for direct access views.

newer, less salient or substantial, or less relevant to our core identity. This can certainly be the case, and a similar kind of response can be provided by a proponent of the ISA theory to explain why some of our self-reports will remain consistent despite perceptual cues, while others are influenced by them.

While we do not want our propositional attitudes to be rigid and unchanging, we want them to change on a rational basis. We want the changes to be for good reasons, not to change randomly, or change due to what we take to be irrelevant or trivial contextual differences. While in some cases having extraneous factors influencing our decisions and judgements might be useful or rational (if there is no basis for picking one set of stockings or another, it will be useful to just pick one anyway) [18], the quality of our handwriting having an influence on our confidence is not the kind of influence that we should rationally want or should expect to have any kind of beyond-the-moment impact. Otherwise, if it were beyond-the-moment, our propositional attitudes would be shifting all the time.

It is notable that most of us are not considered to be volatile, and others who know us well find us largely predictable, even in some cases being better at reading our states than we are ourselves (Schwitzgebel 2008, p.629). Substantial unpredictability in those we know well is a remarkable state of affairs, and often a cause for concern that

[18] In the tale of Buridan's Ass (Zupko 2018), a hungry donkey is equidistant between two identical bales of hay. As there is nothing to make one bale of hay a better option than the other, the donkey is unable to choose between them and after a lengthy period of indecision the donkey dies of hunger. While the donkey had no rational basis to choose between the bales, it *was* rational for the donkey to make a choice anyway. Sometimes making a choice is better than making no choice. Hence, the usefulness of various cognitive heuristics in decision-making, such as responding to ordering effects, which lead to us (for example) preferring the last set of stockings in a line of them, or perhaps the bale of hay on the left. However, we are introspectively blind to these kinds of factors which influence our judgements and so misattribute false explanatory reasons and propositional attitudes to ourselves.

something is 'wrong'. Third-party mindreading may be limited, it may be prone to various errors and mistakes, but in the round it is pretty good. This is not compatible with the 'fragile attitudes' view, because if we were that mercurial, third-party mindreading would be far less reliable than it is. In predicting and responding to others we treat them as rational agents, who change their minds for rational and understandable reasons, not trivial external causes. If our propositional attitudes shifted with such trivial exposure as to shaky handwriting, it would be highly challenging for others to succeed in their mindreading attempts, even if they were aware of the vast multitude of things we may have been exposed to since the last time they saw us. We could become rather different in our outlook and behaviours between breakfast and dinner. This does not match our experience of ourselves and our experience that most people are relatively stable in their attitudes over time. Nor does it match our approach to social interaction.

This consequence gives the wrong picture of our minds, as reflected in our experience of ourselves and the general success of mindreading. While we can be influenced in the now by perceptual cues our experience of ourselves and the success of mindreading suggests that changing our minds tends to be a matter of reasons and internal coherence. Given this mismatch between the implications of accepting Andreotta's view and how relatively stable we find propositional attitudes to be, we should reject Andreotta's objection to the interpretative view of self-knowledge.

Access versus Interpretation

We expect the influence of trivial factors on our propositional attitudes or self-reports of them to have a relatively short-lived influence upon us. This is implicit in the range of psychological studies (including various priming and framing studies) carried out to investigate these kinds of influences; generally the experimenters do not attempt to restore participants to their pre-intervention state via an appropriate debrief[19]. It is not considered unethical to carry out these kinds of studies without some kind of restorative debriefing because the implicit belief is that any effect will be relatively fleeting. We do not expect trivial 'in the moment' perceptual influences to have a lasting effect on our propositional attitudes. Otherwise, imagine the long-term damage you could wreak, by tricking your enemies and competitors into writing things with their non-dominant hand! Our psychological states are more resilient and stable than that. However, this raises a problem for direct access views, namely if our propositional attitudes are changed by trivial extraneous factors, something must happen to change our propositional attitudes back to be more or less the same state as they were before, and it is not clear what this mechanism is or how it occurs.

As an empirically observed phenomenon, this apparent changeability or variance to our propositional attitudes or self-reports—such as that observed in the shaky handwriting experiment—caused by extraneous factors needs to be accounted for by the relevant theory. The interpretation-based explanation is relatively straight-forward; various bits of information, including sensory stimuli, are drawn upon to interpret one

[19] And I am assuming this specifically in the case of the Briñol and Petty (2003) study, as no mention of a debrief or restorative action is reported in the journal article.

as having a given propositional attitude, which is then reported. So, even if we would not expect this particular effect of shaky handwriting leading to lead to a lower level of confidence being reported, we at least have a causal story it can fit into. We might be surprised by this instance of self-reporting a lower confidence due to a minor matter, but not by the general pattern of cause and effect.

Under the interpretation-based explanation, what is being reported is an interpretation of what the individual's relevant propositional attitude might be. There is no reason to suppose that any propositional attitudes have changed due to the influence of perceptual cues, only that the result of the interpretation has been influenced. Thus, if an interpretation-based account is true, the divergence in self-reports due to the shaky handwriting compared to typical self-reports will only last as long as the perceptual cues are salient. Once the perceptual cues are no longer salient, the interpreted self-reports will no longer be influenced by those cues. This is a merit of the interpretative account; it caters for the phenomena of external stimuli influencing our self-reports, while also accommodating the relative stability over time of those self-reports. This is the case, because it is the reports that are influenced by trivial factors not (necessarily) the underlying propositional attitudes. The stability that an interpretation-based account implies matches our experience with ourselves and other people. Yes, there is some variation in our propositional attitudes over time, but we are largely consistent and predictable to those that know us.

In comparison, direct access views do not account for how these various trivial external stimuli lead to differences in our propositional attitudes in the short term while also

retaining relative stability over longer timescales. The view being proposed is that we are not making a kind of mistake when giving our self-reports (there is no confabulation), but it is our propositional attitudes which change based on sensory and contextual information which we then accurately report. The direct access view has to account for how relative stability beyond the immediate short-term arises if our propositional attitudes are changed by trivial external stimuli. Either those changes brought about by trivial factors last well beyond the period where those cues are salient, in which case contra to what we find to be the case we would be mercurial beings who are poor targets for mindreading, or the effect of those trivial factors on our propositional attitudes is short lived. However, while the interpretative view has the resources to give a causal reason why external influences on our self-reports are short-lived, the 'fragile attitudes' view does not. While a direct access account could be shored up to respond to this challenge by adding auxiliary elements to the account, there would remain unresolved issues. One such problem is that if the propositional attitudes themselves are influenced by the perceptual cues it is not clear what would serve as a reference point for a return to a typical pre-influence state.

If our propositional attitudes are fragile and easily influenced by extraneous stimuli, then it casts some doubt on claims that our self-knowledge is especially reliable, which is normally part of direct access views. In cases like that of the shaky handwriting and split-brain patient experiments, for at least some of our propositional attitudes we cannot reliably report what our propositional attitudes are at the point of being asked. In these cases, the propositional attitudes change as part of the response to the question. One could say that this is a good thing, that at the point of employing or

reporting on our propositional attitudes they are in effect judged and updated considering relevant (and as a side effect, irrelevant) information. However, it is not clear whether what is happening is best described in terms of propositional attitudes being updated and then reliably accessed or the generation of an interpretation to report.

In trying to accommodate confabulation data like that found in the shaky handwriting experiment, direct access views appear to need to embrace the idea that our propositional attitudes themselves (and not our self-attributions) are changed by extraneous trivial factors. This leads to the 'fragile attitudes' view, whereby we are mercurial beings. But this view gives the wrong picture of our minds. While some of us may be a little fickle, and we may hold attitudes due to causes and influences that we are not consciously aware of, we are by and large stable and predictable in terms of our propositional attitudes. Generally, while we may be influenced in the short term by contextual factors, longer-term changes to our propositional attitudes are much more rational than the fragile attitudes view implies.

5 Lack of Incongruity Between Self-Reports and Behaviour

A general point raised in Andreotta's objections to Carruthers' claimed cases of confabulations is the lack of signs that people's self-reports are wrong. Their behaviour from the point of giving the self-report is aligned with the content of the self-report. For example, Andreotta notes that in the writeup of the shaky handwriting experiment, there is no mention or hint of the experiment participants behaving or talking as though they thought their confidence ratings were wrong. The thrust of Andreotta's objection is that a close reading of the experiments discussed by Carruthers does not provide evidence for misattributions of propositional attitudes, because there is no incongruity between declared propositional attitudes and subsequent behaviour.

It is true that unless we are speaking insincerely, we go about our lives giving self-reports we endorse as true. Generally, we also do not go around behaving in conflict with our recent declarations ("I love caramel, but loathe pistachio, so… I'll have the pistachio ice cream please!"). Unless there is such conflict, there does not seem to be a reason to hold that we are misreporting our propositional attitudes. We may doubt whether we are right to hold the propositional attitudes that we have (e.g. "Should I really believe/not believe in god?"), and we may have propositional attitudes that are uncertain ("I'm not sure whether I want to visit the beach or not"), but these cases are not the same as misreporting our propositional attitudes. The gist of Andreotta's position here is that we do not intentionally misattribute our propositional attitudes,

and as there is no conflict between our self-reports and our behaviours there is no reason to think we are making misattributions[20].

Within the bounds of the shaky handwriting experiment design, we have limited reason to believe that the participants are *wrong* in reporting their propositional attitude. There is no conflicting information (e.g. to indicate that the participants realise they have misreported their propositional attitude, or that they are acting completely out of keeping with their declared level of confidence) to suggest the content of the self-report should be different, at least, not in the short term or without reflection about their past levels of confidence in their professional qualities. We may say that there may be something amiss with the reported propositional attitudes from those who wrote with shaky handwriting, that they are unlikely to really reflect the participants' normal or historic level of confidence, and in that sense their self-report is wrong, but Andreotta's point is that we do not have a basis to think that they are making a misattribution at that point in time about their current propositional attitude.

5.1 The Kinds of Error We Should Expect from Interpretive Access

In this section, I discuss the nature of evidence that Andreotta is requiring from the confabulation cases, and how it differs from what we should expect to find in the case of everyday confabulations. The point I wish to argue for is that in the case of everyday confabulations, we should not expect to find much in the way of evidence of incongruity between our self-reports and subsequent behaviour. This makes

[20] I am not claiming that there are *never* conflicts of this type, cognitive dissonance (for example) occurs, but generally we take our self-reports to be true and our behaviour (at least broadly) aligns with them.

confabulations harder to detect or notice unless we are searching for them with a mode of investigation that is able to discern which everyday behaviours may be previously unnoticed confabulations. If this is the case, then the kind of stark evidence of confabulations that Andreotta and other critics would wish to see may not be present even though confabulations are occurring. The level of evidence available will be more mundane, and more ambiguous because we are used to accommodating it into our social lives and the way we think about ourselves and each other. The evidence might be there, but unnoticed or unappreciated.

The starting point for my proposal is that the ISA theory seeks to explain how we come to self-report our propositional attitudes; our route to self-knowledge of propositional attitudes. Under the ISA theory we should expect the interpretative route to self-knowledge to be as good as we currently find our route to self-knowledge to be, because if the theory is right then they are one and the same thing. So, our self-knowledge via interpretation should generally be quite good—because our self-attributions *are* generally quite good—and as such we are likely to behave in line with our interpretative self-reports, particularly after publicly declaring them.

A direct access theorist might insist on evidence of errors, given that the presence of errors (confabulations) is the main motivation for skepticism about direct access views, and the main prediction of the ISA theory (Andreotta p.4866). I accept that the level of evidence that Andreotta is seeking is hard to provide, but I do not think it is hard to provide due to confabulations rarely occurring. Rather, evidence of confabulations is not available or generally hard to come by. There are two main reasons for this:

1) our **self-interpretation tends to be *good enough*** even if not fully accurate, and in everyday cases it provides self-attributions that are sufficiently relevant, plausible, and accurate such that talk of errors is not well placed; and
2) we are **generally blind to the errors that do occur**, through some combination of them not being noticeable, being part of our accepted psychological and social lives, or due to the kinds of errors that others make being *expected*, because we and they are using the same interpretative mechanism to attribute propositional attitudes, so we're making the same mistakes.

In cases where there are confabulations, but they are expected confabulations that fit with the way we understand ourselves and others, it might be said that the confabulation errors can *confirm* the mindreading attributions that we make. As a result, these confabulations do not seem surprising, out of place, like a misattribution, or do not seem like a candidate for being an error. The broad point here is that these kinds of everyday confabulations are the warp and weave of everyday life. They're normal, and we're used to them. They are generally beneath the threshold of being noticed, just as we (generally) don't notice how we walk, and the various small inconsistencies in life. Consider the unreliability or variability in weather forecasts, but we tend to only find it remarkable when the forecast is substantially wrong. Or consider the extent to which we rely on our memories, despite being aware that they're fallible.

5.1.1 Self-Interpretation as Good Enough

The first point that I am making with regards to reasonable evidence in favour of confabulations is that our self-interpretation is *good enough*. The ISA theory is an account of how we gain self-knowledge about our propositional attitudes. It is intended as an account of the way we come to have self-knowledge, it is not meant merely as an account of a defective or unreliable route to self-attributions. As we generally do take ourselves to have a decent amount of self-knowledge, we should therefore expect our self-interpretations to be quite good.

Moreover—as our interpretations draw on a range of contextual factors—when there is a misattribution, we should expect the interpreted propositional attitude to be 'good enough' in most cases, even though it is wrong. What I mean by this is that we should expect the confabulated propositional attitudes to be relevant and plausible, and in keeping with our behaviours and psychological history. The kind of misattributions we are likely to make will be relatively inconsequential, and in everyday contexts escape detection as being false without close scrutiny. If our self-reports were frequently substantially wrong, confabulation would not be a strange thing discussed by psychologists and philosophers, it would be a phenomenon we were all well aware of (Cushman 2012, p.350). Additionally, our self-reports would not be helpful s to understand ourselves or each other and would be relatively unsuitable for supporting social interactions.

Note that I am not endorsing a kind of interpretationism, whereby the states you're in are those that you interpret yourself to be in. We can be wrong about what our

propositional attitudes are—we can confabulate them—and behavioural indicators could point clearly to us having a particular propositional attitude, which nevertheless we do not have. The point I am making here is that our self-interpretations do not *need* to be exactly right, they do not need to be highly accurate. They need to be *right or accurate enough*; in the right ballpark, as it were. Close enough to make sense based on the available evidence, is good enough for daily interactions and for everyday confabulations to go unnoticed or be discounted as being down to factors such as vagueness or imprecision. Perhaps when questioned I report that I love philosophy, when actually, I only love a few specific philosophical topics. Or maybe I express regret at what I had done, but really what I regret is the impact I had on others. In these cases, the self-report would likely be *right enough* even if not fully accurate.

Consider, when asked about our confidence in our abilities, like with the shaky handwriting example, just how much accuracy and consistency in our self-reports is needed. Perhaps as I interpret and confabulate my propositional attitudes one day I report that I am rather confident about something, and the next day I report that I am quite confident. We may not notice this difference in our own self-reports or those of others, even if the self-reports came quite close together. Or consider when people provide justification for their election votes and political choices. Many voters when asked why they voted a certain way will report with conviction on their beliefs or endorsements on various talking points and political messaging, despite this being an unlikely causal factor in their actions (Murphy-Hollies 2022, pp.111-116).

It may only be in specific contexts where there is a higher standard for accuracy and precision (such as scrutiny of experimental participants by psychologists and philosophers) where discrepancies or inaccuracies in our self-reports are noticed and make a non-trivial difference. In other situations, they may be less noticeable, or easier to explain away. Scrutiny of my self-reports is likely to be higher in a philosophical debate, psychology experiment, or therapy session than they are in ordinary conversation. In these cases my declared propositional attitudes may need defending or explaining, or I might be put in a position to test how specific my attitudes are.

The evidence of people not behaving in direct conflict with their self-reports will be rare, but this does not count strongly against there being no confabulations. The ISA theory predicts everyday confabulations, of misattributing propositional states to ourselves just as we do when it comes to others (Carruthers 2013, p.6). This amounts to some proportion of our self-attributions being wrong, either entirely or, most commonly, to a degree. This is certainly the case when it comes to interpreting others. We are not as accurate as we would like, but quite often what we attribute to others is close enough to be good enough for our purposes.

As interpretative self-attribution will generally be right, or right enough, most of the time, in those cases where there is a misattribution, generally we will act in congruence with the misattribution, because the gap between our actual propositional attitudes and what we self-report will be relatively small. For example, If I report that I *adore* the film Palm Springs (say an equivalent of a 9 out of 10), it is not the kind of thing that readily yields to a criticism that *actually* I only *love* the film (say an 8.5 out of 10). Either

answer, or a similar set of answers, are perfectly sufficient. They are accurate enough, useful enough, and unlikely to raise suspicion of error if I give either one, even though one might be a 'true' indication of my propositional attitude, and the other a misattribution.

Similarly, I might make misattributions that nevertheless seem to fit with my overall behaviour and other self-reports. This is likely to happen, because according to the ISA theory we're drawing on a range of contextual information, like our declared intentions and interests, and the activities we're currently involved in, so we should expect our self-reports to have a degree of congruence with such things. For example, after helping a friend with household chores I might report that I did it because I wanted to be helpful, and this might fit with my general intentions and track record of behaviour, but in this specific example I might have helped more out of a sense of obligation, which I had forgotten after completing the task and felt satisfied with what I had achieved.

While the outcome of our normally 'good enough' mindreading of third parties could be readily falsified via feedback that contradicts the mindreading results (at first Smith believed that Jones liked the painting, but Jones said that he did not, so Smith revised his belief), the same is not true for our self-knowledge (Schwitzgebel 2008). We normally do not encounter contradictions to the results of our attempts at self-knowledge, because there are limited alternative means of finding out about our own mental states, as well as our self-reports being 'good enough' that there isn't much opportunity to find clear counterevidence. We are restricted to a limited (and

potentially misleading) perspective, with any contradicting feedback largely arising from the outside. When we do encounter contradictions to our supposed self-knowledge, such as conflicting behaviour, they must compete against the experiential weight of a lifetime of having a sense of what our mental states are that is rarely contradicted, and such contradictions can be easily explained away ("I changed my mind").

A further issue is that in everyday cases when we misreport our propositional attitudes, if we get them wrong, it might be discounted as being a case of changing our minds, or a matter of imprecision in expression, rather than the propositional attitude being incorrectly attributed at the time. If Smith used to say his favourite colour was red, and now says that it is blue, we might not know whether this has been a genuine change of preference, or some kind of error. While it does happen, people rarely tell us we are wrong about our own mental experiences (Schwitzgebel 2008), at least not as adults. They have little in the way of epistemic warrant to challenge us; often they have no way of knowing when we are wrong. They may say something like "You don't really believe *that*, do you?", but what is usually meant by this is "you should change your mind, rather than continue to believe this foolish thing", rather than "you just made a mistake about what you believe". So, all in all, we have very limited feedback when we are wrong about what our propositional attitudes are, whether the error is detected by ourselves or by others.

5.1.2 Blind to Error

The second point which I have raised as to why evidence of confabulations is difficult to detect or notice, is that we are generally blind to the self-attribution errors that do occur. We are unlikely to notice such everyday confabulations in others or ourselves in part because we are *used* to them. They are likely part of our everyday social life. We do not notice these confabulations because they are routine, banal, or below the threshold of our notice, like believing we chose something based on its inherent qualities rather than factors like ordering and priming effects (e.g. Nisbett and Wilson 1977).

Everyday confabulations are likely commonplace and just go unnoticed. They are given, just as with accurate self-reports, to explain people's behaviour to themselves and to others, to signal to others that they are rational and trustworthy social agents, and to give a positive self-image (Ganapini 2019, pp.196-197; Murphy-Hollies 2022, p.114; and Sullivan-Bissett 2015, p.552). We are prompted to give an answer, we give one, and generally the confabulations will seem plausible and to fit with the available evidence and context. This is because confabulations are not a case of our cognitive machinery being broken or going wrong. Rather, what is happening is a general everyday case of giving a self-report, interpreting ourselves, drawing upon the available evidence, and giving a plausible and coherent answer, which sometimes will be a correct attribution, and sometimes will be a misattribution.

In addition to being used to everyday confabulations, there are no markers for confabulation for us to detect. The experience of a correct self-attribution and a self-

misattribution is the same. There is not a sense of wrongness or lack of confidence in a misattribution to distinguish it from a correct self-attribution[21]. Unless our misattributions are substantially and obviously wrong, we have little to no way of falsifying them. We are blind to much of the workings of our mind, so we cannot 'look into the box' and discover that we are wrong, and even if we do notice something—like an apparent change in our preferences—we likely discount them on the basis that we have changed our mind. And in many cases of other people confabulating, if we are using the same capacity to mindread them, we may come up with the same misattribution as they do (or at least find it to be relevant and plausible) and so their (false) self-report provides confirmation of what we already believed.

When we confabulate, like when we erroneously say that we intend to get a Coke—based on factors such as being thirsty, seeing Coke, liking Coke, believing we could obtain Coke—not only are we likely still reporting a propositional attitude that is relevant and plausible, once we declare our self-report we are reinforcing the misattribution and making a public avowal of our intention. There is relatively little room to detect a misattribution, particularly when they can become self-fulfilling. If someone were to tell us we had not really intended to get a Coke, absent some conflicting information from another self-report, it is not clear why we would choose to believe them. We're thirsty, we like Coke, here is a Coke that belongs to us, and so we

[21] Some of our self-reports may contain uncertainty or a lack of confidence: I don't know if I like bubblegum flavoured ice cream or not; I am not confident whether I want to go hiking or not; I am sitting on the fence as to whether I believe political party A will handle the economy better than political party B. But that is a feature of the content of the self-report; I do not lack confidence in the self-report itself. While it may be a matter of my perspective and experience, it seems like *felt confidence* in a propositional attitude (as opposed to confidence based on reflection, investigation, or analysis) is not even a thing. The self-report just-is, and 'of course' it is right.

drank some, so of course we think we had an intention to drink the Coke. Mere third-party testimony that we didn't have that intention has little weight against the competing reasons to believe otherwise. Yet, it is entirely possible that we hadn't really intended to get the Coke, and we absent-mindedly picked it up and drank some, or we did so through habit, or someone switched drinks on us from a Pepsi to a Coke and we didn't really notice[22]. For there to be a clear recognition of a misattribution it would need something like a declaration that we want a Coke, followed by a realisation that we don't want a Coke, leaving some confusion as to why we gave a false self-report rather than thinking we had changed our mind.

When we are asked questions about our everyday decisions and actions, we can usually give a prompt response. "I'm angry because Jones is plotting against me". "I'm going to get a Coke". "I prefer this wine to the other wine because it is more pleasant". These seem like natural and true responses, but they may not be true. Even for something as simple about why we prefer a particular wine, we may be wrong. Plassmann *et al*. (2008) demonstrated that people will rate a wine as tasting more pleasant when the label is switched to have a higher price. These people are reporting in good faith on a difference in taste experiences but in situations like this people do not say things taste better *because* they're more expensive, they offer up confabulatory judgements. Chartrand and Bargh's (1999) research showed that people tend to unconsciously mimic the behaviour of other people (in this case strangers), and that

[22] We know from various studies—such as an experiment by Simons and Levin (1998), whereby the person which participants were talking to was swapped while their view was blocked by two people carrying a door between them—that we are sometimes *change blind*, and act after the change as though no change had occurred, and would when prompted give misattributed propositional attitudes with regards to the situation to give a *post hoc* rationalisation.

mimicry can increase liking between interaction partners. Yet, if people uninformed of this type of psychological research are asked questions like “why did you cross your legs during the conversation?” and “what made you like this person?” it is unlikely that they would give mimicry as a reason, even though that is a contributing reason. These are simple examples but show how commonplace confabulations are when it comes to our self-knowledge. It’s not that we just do not know the answer to questions like “why did you cross your legs during the conversation?”, it is that we don’t know the answer *and* make something up and think it is as true as all the other self-reports we give. If people confidently provide answers they believe to be true about their behaviour that do not relate to the actual underlying reasons for that behaviour, then the answer they give is fabricated and is a confabulation. But we do not recognise that there has been a confabulation. In the context of everyday interactions we have no real way of knowing in our own case or in third-party cases whether a self-report is a confabulation, because everyday confabulations are plausible and will fit the evidence in a good-enough way. Confabulations are part and parcel of our everyday experience, but they are such an everyday part of our cognition and experience that they go unnoticed.

6 Summary

In this paper I have defended the ISA theory by arguing against important objections raised by Andreotta (2021). First, I have emphasised the need to look for evidence for misattribution (i.e. confabulations) retrospectively from the point of provoking a self-report, because that is the direction in which the error generally lies, but also such a tactic sidesteps any confounding issues of self-fulfilment following a self-report. Secondly, I have argued that the ISA theory better accounts for empirical data relating to trivial extraneous perceptual cues influencing the content of our self-reports, in contrast to Andreotta's proposal that our propositional attitudes are fragile and are changed by trivial perceptual cues. Adopting the ISA theory view avoids an unsatisfactory counterfactual outcome of Andreotta's view, that we would be mercurial and far more difficult to mindread, which is in conflict with how useful we find mindreading to be. Finally, I have sketched a picture of what kind of evidence we can plausibly expect to find for everyday confabulation, given the nature of our social interactions and the likelihood of us being blind to or overlooking non-jarring confabulations. Taken in combination, these lines of argument make important progress in defending and supporting Carruthers' ISA theory of self-knowledge. In doing so, they also show that the ISA theory provides a plausible and productive account of self-knowledge: it offers an explanation for the way in which our reported propositional attitudes are affected by perceptual cues and other factors and it offers an explanation of the similarities in performance between self-attribution and third-party mindreading. Overall, the ISA theory seems the more plausible account for self-knowledge and better able to accommodate the confabulation data.

Finally, I wish to make a broad point. The prevailing default assumption is that we do not confabulate. Confabulation is not something that fits easily into folk psychology, and most lay people would likely be astonished by the evidence of confabulation emerging from psychology experiments. Accepting that we confabulate is counterintuitive to the way we think about ourselves and how our minds work. We might even be predisposed to be resistant to the idea that we confabulate (or even that we self-interpret our propositional attitudes). Finding evidence for everyday confabulation is not just a matter of pointing to cases of confabulation, it is a matter of convincing people to work against their intuitions and accept that in some cases people are making mistakes about their own mind, even though it might not seem that way to them or to others.

7 References

Andreotta, A. (2021) 'Confabulation does not undermine introspection for propositional attitudes', *Synthese*, 198, pp.4851-4872.

Balsvik, E. (2017). Interpretivism, first-person authority, and confabulation. *Philosophy of the Social Sciences*, 47, 311–329.

Bayne, T. and Spener, M. (2010) 'Introspective Humility', *Philosphical Issues*, 20, 1, pp.1-22.

Briñol, P. and Petty, R. (2003) 'Overt Head Movements and Persuasion: A Self-Validation Analysis', *Journal of Personality and Social Psychology*, Vol. 84, No.6, pp.1123-1139.

Byrne, A. (2005) 'Introspection', in Chalmers, D. (Ed.) (2021) *Philosophy of Mind: Classical and Contemporary Readings,* pp.615-626. Oxford University Press.

Byrne, A. (2012) 'The Opacity of Mind: An Integrative Theory of Self-Knowledge', *Notre Dame Philosophical Reviews*. Accessed 30/06/2023, available at: https://ndpr.nd.edu/reviews/the-opacity-of-mind-an-integrative-theory-of-self-knowledge/

Butler, J. (2013) *Rethinking Introspection*. Palgrave Macmillian.

Carruthers, P. (2009) 'How We Know Our Own Minds: The Relationship Between Mindreading and Metacognition', *Behavioural and Brain Sciences*, 32, pp.121-182.

Carruthers, P. (2013) *The Opacity of Mind*. Oxford University Press.

Carruthers, P. (2018) 'There is no such thing as conscious thought', interview by Ayan, S. in *Scientific American*. Accessed 04/06/2023, available at: https://www.scientificamerican.com/article/there-is-no-such-thing-as-conscious-thought/

Chartrand, T. and Bargh, J. (1999) 'The Chameleon Effect: The Perception-Behaviour Link and Social Interaction', *Journal of Personality and Social Psychology*, 76(5), pp.893-910.

Cushman, F. (2012) 'Understanding Confabulation', in Brockman, J. (ed.) *This Will Make You Smarter*. Harper Perennial.

Dennett, D. (1993) *Consciousness Explained*. Penguin.

Doyle, R. and Voyer, D. (2016) 'Stereotype manipulation effects on math and spatial test performance: a meta-analysis', *Learning and Individual Differences*, Vol. 47, pp.103-116.

Dutton, D. and Aron, A. (1974) 'Some Evidence for Heightened Sexual Attraction Under Conditions of High Anxiety', *Journal of Personality and Social Psychology*, 30(4), pp.510-517.

Engelbert, M. and Carruthers, P. (2010) 'Introspection', *WIREs Cognitive Science*, 1, pp.245-253.

Fricke, M. (2014) 'Transparency of Opacity of Mind?', *Analytical and Continental Philosophy: Methods and Perspectives*, *37th International Wittgenstein Symposium*, pp.97-99.

Ganapini, M. (2019) 'Confabulating Reasons', *Topoi*, 39, pp.189-201.

Gazzaniga, M. (2011) *Who's in Charge? Free will and the science of the brain*. New York: Harper Collins.

Hutto, D. and Ravenscroft, I. (2021) 'Folk Psychology as Theory', in Zalta, E. (ed.) *The Stanford Encyclopaedia of Philosophy* (Fall 2021 Edition). Accessed 02/06/2021, available at: https://plato.stanford.edu/entries/folkpsych-theory/

Kind, A. (2013) 'The Opacity of Mind: An Integrative Theory of Self-Knowledge', *Analysis Reviews*, Vol.7, No.1, pp.172-174.

Kind, A. (2021) 'Introspection', *The Internet Encyclopaedia of Philosophy,* ISSN 2161-0002, Accessed 05/11/2016, available at: http://www.iep.utm.edu/introspe/

Linder, D., Cooper, J. and Jones, E. (1967) 'Decision Freedom as a Determinant of the Role of Incentive Magnitude in Attitude Change', *Journal of Personality and Psychology,* 6(3), pp.245-254.

Luca, B. and Gordon, R. (2017) 'Folk Psychology as Mental Simulation', in Zalta, E. (ed.) *The Stanford Encyclopaedia of Philosophy* (Summer 2017 Edition). Accessed 02/06/2024, available at: https://plato.stanford.edu/entries/folkpsych-simulation/

Lyons, W. (1986) *The Disappearance of Introspection*. The MIT Press.

Mandik, P. (2014) *This is Philosophy of Mind*. Wiley Blackwell.

Marraffa, M. (2021) 'Theory of Mind', *The Internet Encyclopaedia of Philosophy,* ISSN 2161-0002/. Accessed 02/07/2023, available at: https://iep.utm.edu/theomind/

Murphy-Hollies, K. (2022) 'Self-Regulation and Political Confabulation', *Royal Institute of Philosophy Supplement*, 92, pp.111-128.

Nisbett, R., and Wilson, T. (1977) 'Telling More Than We Can Know: Verbal Reports on Mental Processes', *Psychological Review*, Vol.84, Number 3.

Plassmann, H., O'Doherty, J., Shiv, B. and Rangel, A. (2008) 'Marketing actions can modulate neural representations of experienced pleasantness', *Proceedings of the National Academy of Sciences (USA)*, 105(3), pp.1050-1054.

Rimkevičius, P. (2020) 'The Interpretive-Sensory Access Theory of Self-Knowledge: Empirical Adequacy and Scientific Fruitfulness', *Problemos*, 97, pp.150-163.

Ryle, G. (2000) *The Concept of Mind*, Penguin Classics.

Scaife, R. (2014) 'A Problem for Self-Knowledge: The Implications of Taking Confabulation Seriously', *Acta Analytica*, 29 (4), pp.469-485.

Schroeder, T. (2006) 'Propositional Attitudes', in *Philosophy Compass*, Vol. 1, Issue 1, pp.65-73.

Schwitzgebel, E. (2008) 'The Unreliability of Naïve Introspection', in Chalmers, D. (Ed.) (2021) *Philosophy of Mind: Classical and Contemporary Readings*, pp.626-641. Oxford University Press.

Schwitzgebel, E. (2024) 'Introspection', in Zalta, E, (ed) *The Stanford Encyclopaedia of Philosophy* (Summer 2024 Edition). Accessed 02/06/2024, available at: https://plato.stanford.edu/archives/win2019/entries/introspection/

Sellars, W. (1956) 'Empiricism and the Philosophy of Mind', in Chalmers, D. (Ed.) (2021*) Philosophy of Mind: Classical and Contemporary Readings*, pp.415-422. Oxford University Press.

Simons, D. and Levin, D. (1998) 'Failure to Detect Changes to People During a Real-World Interaction', *Psychometric Bulleting & Review*, Vol. 5, No.4, pp.644-649.

Smithies, D. and Stoljar, D. (2012) 'Introspection and Consciousness: An Overview', in Smithies, D. and Stoljar D. (Eds.) *Introspection and Consciousness*, pp.3-28. Oxford University Press.

Spener, M. (2013) 'Moderate Scepticism About Introspection', *Philosophical Studies*, 165, pp.11847-1194.

Spener, M. (2024) *Introspection: first-person access in science and agency*. Oxford University Press.

Sullivan-Bissett, E. (2015) 'Implicit bias, confabulation, and epistemic innocence', *Consciousness and Cognition*, Vol. 33, pp548-560.

Wells, G. and Petty, R. (1980) 'The Effects of Overt Head Movements on Persuasion: Compatibility and Incompatibility of Responses', *Basic and Applied Social Psychology*, 1(3), pp.219-230.

Williamson, T. (2020) *Philosophical Method: A Very Short Introduction*. Oxford University Press.

Zupko, J. (2018) 'John Buridan', in Zalta, E. (ed) *The Stanford Encyclopaedia of Philosophy* (Fall 2018 Edition). Accessed 13/11/2023, available at: https://plato.stanford.edu/entries/buridan/

USEFUL BUT NOT ACCURATE: AN ARGUMENT FOR ILLUSIONISM

We experience things like redness *and* painfulness*—the qualitative aspects of experience—which seem difficult to account for in terms of the brain and information processing. This difficulty arises, because the qualitative aspects of experience seem to have features which (when taken at face value) seemingly do not fit comfortably with a physicalist and mechanistic view of the brain. There are several different approaches to resolving this difficulty, including Keith Frankish's (2017; 2023) Illusionism. According to Illusionism we only seem to experience things like* redness *and* painfulness*. They do not really exist, rather they are the result of a misrepresentation or illusory perspective, somewhat analogous to the apparent existence of rainbows as arcs in the sky.*

In this paper I present a novel argument in support of Illusionism, that as evolved entities we should expect our various mental capacities to track usefulness rather than accuracy. As such the way things seem to us seem the way that they do because they have developed to be useful, rather than truth-tracking, and so we should not expect our qualitive experiences to provide us with a view of how things really are. The various perceptual illusions and cognitive biases that we experience show that in a range of cases such expectations are correct. As a result, we should be wary of using our

qualitative experiences as a basis for determining what things really are, and in particular the metaphysical status of those experiences.

1 Introduction

Imagine a robot. It can find red apples. When you kick it in the leg it hops about and complains.

Can the robot genuinely see the apples as we do? Can it experience the *redness* of the apples? Does it experience a *painful* sensation where you kicked it in its leg? Intuitively, and as some data suggests (Diaz 2021), most people would answer these questions with a resounding 'no'.

Now, imagine a future robot. It is a duplicate of a human being but composed of microchips, servos, and some cunning new materials. This robot looks and behaves exactly like a human being. It smiles. It laughs. It occasionally bashes its toes on something and screams. It will tell you about its holiday and express a love of caramel flavoured ice cream. You cannot tell it apart from 'genuine' human beings. Can this robot really see the apple as we do? Can it experience the *redness* of the apple? Does it feel *pain* when you kick it? While some (e.g. functionalists) might say yes, many others will still say no. But why answer 'no'?

A common view is that that conscious experience is something leftover when all the functional goings on of the mind are accounted for: it is something extra (Tye 2021). The idea is that there is something it is like to be 'us' and to have experiences (Tye 2011) which is lacking in robots, and more obviously lacking in things like rocks and

rivers. And so, the reasoning goes, because conscious experience is not involved in any functional goings on, consciousness is not going to be explained by science—at least not science as we know it—anytime soon (Chalmers 1995, pp.93-95). According to this view, scientists will not be able to work out what 'phenomenal properties' like *redness* and *painfulness* are—or why they appear a certain way—by looking at neurons and brain structures in ever increasing detail. Indeed, to many, our conscious experience is incompatible with a physicalist worldview.

This problem is an explanatory gap problem, and explaining what consciousness is and where it comes from in terms of causes and grounding is known as the Hard Problem of Consciousness (Chalmers 1995). However, not everyone accepts there is a Hard Problem (e.g. Churchland 1996; Dennett 2021; Frankish 2017; and Strawson 2018). One response to the Hard Problem is to argue that the various intuitions and introspective judgements which lead to the belief that there is a Hard Problem arise from the illusory nature of phenomenal properties (Frankish 2017 and Frankish 2023). This position is *Illusionism*.

In this paper, I present a novel argument in support of illusionism that we should not rely on introspection to determine the true nature of our experiences, and specifically not the nature of *feels* (phenomenal properties) like *redness* and *painfulness*. The reason we should not rely on introspection for this purpose is twofold. First, we should not expect introspection to be fit for this purpose due to it being an evolved psychological process, and as such we should expect it to have evolved to track usefulness rather than truth or accuracy. Secondly, we find no reason to believe

introspection is fit for determining the true nature of our experiences, drawing analogously on perceptual illusions and cognitive biases. While Frankish (2017 and 2023) describes the illusory nature of the appearance of phenomenal properties as resulting from misrepresentations, I want to emphasise that these misrepresentations are useful rather than maladaptive.

In the following sections, I introduce the Hard Problem of Consciousness and then present the case for Illusionism. Then, I argue that the appearance of experience is useful rather than accurate and give a defence in response to some of the stronger criticisms of Illusionism which could be levelled against my argument.

2 The Hard Problem of Consciousness

2.1 Qualitative Experience and Physicalism

To be conscious is to be aware of something. This awareness can manifest in various forms and encompass different types of things. It includes the experience of perceptual content and mental states such as sensory and somatic experiences, conscious imagery, emotions, thinking, desiring, and dreams (Schwitzgebel 2017, pp.227-229). We can be aware of external objects and sounds, such as seeing a tree or hearing music. We perceive the vibrant *redness* of an apple, feel the stinging *painfulness* of being kicked in the shin, and experience the creeping *dread* of a monster under the bed.

Subjective experience is notoriously difficult to define and is often approached through referring to shared experiences or the notion of 'what it is like' (Nagel 1974; Schwitzgebel 2017). Imagine looking at a red apple beside a green apple. When you see the *redness* of the red apple and the *greenness* of the green apple, you have an awareness of their colours that transcends mere acknowledgements of difference. We have this sense of their different colours; there is something *it is like* to see them. We do not just get a sense that something is red or green, as though something in our brains has reached a conclusion and reported on it as some dry fact. Their *redness* and *greenness* seem to be right there, apparent to us.

Similarly, there is something *it is like* to put your hand into a bucket of ice, something more than simply registering that the temperature is lower. You may experience the

initial shock, the unpleasantness, the *coldness*, and the feeling of resistance of the ice. These aspects are part of what it is like for you to plunge your hand into the depths of an icy bucket. Registering the fact that the apple is red or that your hand in the bucket of ice is cold is different to the experience of seeing the *redness* or feeling the *coldness*. So too, when we are kicked in the shin, there is a *painfulness* to it that seems to be something different to a mere registering of the impact. These experiences seem real: it seems like the apple has a vivid *redness* and there is a throbbing *painfulness* after being kicked in the leg.

This what-it-is-like approach to picking out conscious experiences does not carry much explicit content or provide much of an explanation or definition of the experienced character of consciousness. Instead, it merely points to instances of what is meant, so we get an idea of the kind of phenomenon that is being discussed, but we do not learn much about the phenomenon itself via our attempts to describe it. Instead, we are left to consider for ourselves via introspection what is meant when someone refers to a (apparently) similar example or shared experience. However, we have little sense of how experiences may vary between us. Perhaps I see colours differently to you—maybe your *redness* is more vivid and vibrant than mine—or perhaps if I could somehow directly experience what it is like for you to feel pain, I might find that we each feel *painfulness* differently.

The subjective character of experience—like *redness* and *painfulness*—is often considered to be puzzling and (at present at least) difficult to account for in physicalist terms. This subjective character is sometimes thought to be something in addition to

the representational or functional role of what we are conscious of. Ned Block (1995) distinguishes between two types of consciousness: access consciousness and phenomenal consciousness. Access consciousness refers to the aspects of our mental life that we can report on, reason about, and use to guide our behaviour. In contrast, phenomenal consciousness refers to the qualitative, what-it-is-like aspects of our experiences. Indeed, it is thought that what makes these properties phenomenal is that there is something it is like to have them (Hall 2021, p.11002). According to Block, these two types of consciousness can exist independently of each other, i.e. the phenomenal aspect can be separate from the accessible content. The idea is that absent the *painfulness* of pain, if someone kicked us in the leg we might know that we have suffered damage, we might consider that to be bad, we might want to avoid it happening again in future, and information about the event might be prioritised such that we find it difficult not to attend to it, but it would not be *painful*.

This view aligns with phenomenal realism, which posits that the *feels* of experience—the phenomenal properties—are irreducible to physical processes or functional roles (Chalmers 1995; Frankish 2023; Nagel 1974; Sahu 2022). Others, including functionalists, reject this view, arguing that the *feels* of experience can be fully explained by its representational and functional roles. This is perhaps the key disagreement in whether to accept phenomenal realism: whether there is something distinct from the physical and functional properties of an experience that gives rise to how it feels (Niikawa 2021, pp.12-14).

As our science and understanding of the brain improves, there seems to be less room for the phenomenal to fit in. The qualitative content of experience seems to be increasingly difficult to reconcile with what we see in terms of brain activity. The brain is made up physical-stuff with physical goings-on, such as networks of neurons with signals travelling across them. How, the worry goes, can these networks of firing neurons lead to the experience of *redness*? We may be able to track how the brain detects and discriminates red in what we see, but beyond an informational state that stands-for certain wavelengths of light, how do we end up with an experience of *redness*? Or of other experiences like *painfulness* or *dread*? We can imagine a robot responding as though it were in pain when we kick it in the leg, but amongst all the information states and its behavioural responses, where can the *painfulness* of the event be found? This is not a new kind of worry, in 1714 Leibniz asked us to imagine:

> "… that there is a machine whose structure makes it think, sense, and have perceptions, we could conceive it enlarged, keeping the same proportions, so that we could enter into it, as one enters into a mill. Assuming that, when inspecting its interior, we will only find parts that push one another, and we will never find anything to explain a perception."

While science has made great progress in understanding the mechanisms that give rise to perception, the problem remains how those mechanisms—brain processes—give rise to perceptual experiences. Illusionism avoids this difficulty by positing that experiences are a kind of judgement or introspective illusion (Frankish 2023 and Shabasson 2022, p.427). The qualitative aspect of experience is not the only challenge

to fitting conscious experience into a physicalist worldview however, and prominent amongst those other challenge are the claimed second-order properties of experiences.

2.2 Second-Order Properties of Experiences

A significant challenge for physicalist worldviews regarding conscious experience lies in the apparent introspectable second-order properties of phenomenal properties that constitute the character of experiences (i.e. the properties of the phenomenal properties). These properties are purported to describe the way phenomenal properties (e.g. *redness* and *painfulness*) are presented to us—or what they seem like to us—akin to the characteristics of a communication medium, much like the presentation format of a film or a piece of music. Frankish (2017 p.15) characterises these second-order properties of phenomenal properties as:[23]

- simple;
- ineffable;
- intrinsic;
- private; and
- immediately apprehended.

What these second-order properties amount to is that the phenomenal properties of consciousness are meant to be, or are considered by some to be: irreducible; indescribable; have at least some element to them that are not relational,

[23] Lists of phenomenal properties can vary between philosophers, but they tend to have some overlap. For example, Chalmers (2018, p.49) lists: intrinsic, non-physical, non-representational, primitive (simple), ineffable, and non-functional.

dispositional, or functional; cannot be directly accessed by anyone else; and in some sense are immediately and intimately known to us (Dennett 1988 and Dennett 2013, p298).

As articulated, this set of second-order properties makes phenomenal consciousness special in a way that some consider to be incompatible with physicalism. So, for example, when we see a bright red apple, the properties of that experience like the *redness* of it are said to be indescribable. The indescribability is not meant to be due to a shortfall in language, like some lack in our vocabulary, or due to the limits of our cognitive abilities. Rather, the indescribability is supposed to be due to something special about the experience itself that means it cannot be expressed linguistically. This means that we cannot adequately capture and communicate what it is like to see the *redness* of the apple such that someone who had not experienced *redness* would appreciate what the experience would be like. Instead, when we talk of experiences, we need to resort to using analogies with other experiences, like comparing the colour of the apple to the colour of a tomato. We point to similar experiences and say, “it is like that”.

The worry is that these second-order properties of phenomenal properties (singularly or collectively) are not compatible with the view that our conscious experiences (and more broadly, our minds) are part of, and only part of, the physical world. Yet despite this apparent incompatibility with physicalism, it is difficult to outright reject the reality of these second-order properties that are part of experiences, as we clearly do have experiences. We can feel the shocking *coldness* when we plunge our hand into a bucket

of ice. We have a feeling of *dread* when considering the monster under our bed. We really do have an experience of *redness* when looking at the apple. If we acknowledge these kinds of experiences, there is pressure to admit that these experiences do seem to have at least some of the second-order properties sometimes ascribed to them (as per Frankish 2017). For example, we really do seem unable to describe the *redness* of the apple in terms that match the experience[24].

How these experiences can be the way they are, given what we know about the brain and the physicalist worldview needs explaining.

2.3 The Hard Problem of Consciousness

There appears to be no satisfying explanation of how we can have experiences with *feels* such as *redness* and *painfulness* grounded in a physicalist worldview. It is widely held that there is an explanatory gap, with brains and neuronal activity on one side, and the *feels* of experience on the other. We know a great deal about the brain, but it seems like brains—three-pound lumps of matter, containing a vast network of neurons signalling each other—are not the sorts of things that we should expect to lead to the *dread* of the monster under our bed or the *painfulness* of being kicked in the shin. This apparent explanatory gap between how things seem is what motivates anti-physicalist views of how conscious experience fits within the world.

[24] Colours can be described in a variety of ways, for example using Red-Green-Blue (RGB) codes. However, the idea is that if you had never seen red before, the various ways that we have of describing or referring to red (like an RGB code) would not impart an idea of what it would be like to see red. You would not know what a new colour was like until you had experienced it.

2.3.1 The Explanatory Gap

We can imagine sophisticated robots created in a factory with electronic brains made to be just like ours. Many of us can imagine such robots talking and participating in everyday life. We can even imagine them yelping and hopping about when we kick them in the shin. But it seems difficult to accept that the robots could feel *pain* just like we do. This suggests that there is some kind of gap at play, with one assembly of physical stuff (a human) feeling 'real *pain*' but with a similar or even functionally identical assembly of physical stuff (a sophisticated robot) not feeling any *pain*.

This apparent gap, where robots perhaps do not feel *pain* despite having all the structure, functions, internal states, and behaviour in place is a type of explanatory gap (Levine 1983). Such gaps do not exist when our understanding of a phenomenon, like heat, is sufficiently explained by understood causal mechanisms, such as the motion of molecules[25]. To many, this does not seem to be the case or even possible with phenomenal consciousness. We can give a causal account for the function of pain, with such an explanation likely to include signalling of damage to the body, aversion behaviours, and so on, but such an account leaves out the *painfulness*. It hurts! Though we can identify physical correlates of pain, it is not clear how they lead to the feeling of pain, or even how *pain* could come about.

[25] Setting aside it may provide a full causal explanation, I remain unconvinced that the explanation of heat in terms of molecular motion offers an intuitive or psychologically compelling bridge between the appearance and reality of heat.

It is important to note that Levine's argument for an explanatory gap for consciousness is one of intelligibility (Levine 1983). It does not establish that there is an unbridgeable divide between the *feels* of experience and a physicalist worldview, rather it highlights that our understanding of the physical underpinnings of experience—neurons firing and the like—does not seem to give a satisfying explanation of phenomena like *painfulness*. What might matter here is what type of gap is present; whether it is a gap of understanding, of not yet having tracked down the relevant causal mechanisms or grounding, one that exists because our theoretical commitments make it unbridgeable, or simply one that we may never find a sufficiently satisfying answer to bridge it with. It may simply be that our perception of there being a gap is due to cognitive limitations, rather than some insight about the relationship between mind and body (Bayne 2022, p.140 and Churchland 1996). But in any case, we are left with an explanatory gap for why certain brain activities are accompanied by a conscious experience, like of *pain*.

2.3.2 The Hard Problem

Bridging the explanatory gap between brains and the *feels* of experience is sometimes known as the 'Hard Problem' of consciousness (Chalmers 1995, pp.92-93). While the brain sciences seem on course to one day explain how the brain works to carry out its various functions—adding numbers, detecting food, coming up with strategies to set monster traps—some argue that the same does not seem to be true for phenomenal consciousness. Chalmers, and many others, are of the view that once the various behavioural and functional doings of the brain are accounted for, phenomenal consciousness would remain unexplained. The key point here is that the hard problem

arises based on the view that phenomenal consciousness does not have a functional role (Chalmers 2018, p.50): the idea that phenomenal consciousness is not *for* anything in terms of things such as behaviour, competence, or processing.

One could argue that we have reason to believe that phenomenal consciousness does have a functional role[26]. Pain seems, well… bad. It is unpleasant. A thing to be avoided. Joy feels good, and something to seek. Hunger is bad too, and it does seem to be situated in parts of the body where food will end up, like signalling that we have a hole that needs filling. When we are thirsty, imagining the cool refreshing effect of some chilled water is rather motivating. Raw *feels* are akin to the bedrock of what we attempt to gesture at when we talk of the basics of human experience.

Pain *feels* bad. We do not like it. We generally seek to avoid it. Those who do not experience pain as bad are likely to be less well adapted to thrive than others, not avoiding or even actively seeking out pain which could lead to debilitating damage to their bodies. If someone has *pleasurable* experiences in place of *painful* experiences, then they would not be in pain. Imagine if extreme debilitating hunger felt *really good*, such that people sought to attain and maintain the experience. This would not be conducive to people being healthy. This suggests that *feels*, whatever they actually are, have a functional role.

[26] Humphrey (2012), for example, argues that the way our mental processes work produces a "magic show" which gives experiences and our sense of the world an 'enchantment'. This enchantment gives the possessor a better and more involved interest in their continued existence and evolutionary/reproductive fitness.

A further point against the idea that phenomenal consciousness isn't *for* anything is that thousands of pages of philosophy have been devoted to phenomenal consciousness, and we are confident we have it, whatever it happens to be. This suggests that phenomenal consciousness has a causal effect on our thinking and has some kind of functional role[27].

Against such a position, the Hard Problemist can still claim there is an explanatory gap even while allowing that phenomenal consciousness has some kind of functional behavioural role to play. One can still raise the issue that pain is *painful*, and that when the various functional goings on are explained, the *painfulness* is something extra and, in a sense, unneeded in addition to our various behavioural responses and attitudes towards *painfulness*-inducing events[28]. At issue here is whether there is something beyond the functional psychological properties of the *feels*, and further whether that something extra is non-physical, or at odds with our current physicalist worldview, versus just appearing to be so (Niikawa 2021).

Explaining phenomenal consciousness is known as the Hard Problem because not only do the *feels* of experience seem unexplainable in terms of the causal mechanisms of the brain that scientists study, but it also seems as though such investigations are not even adopting a workable strategy that could lead to an answer. Mechanism does not seem like the kind of thing to give rise to feelings[29]. According to Chalmers (1996),

[27] This view can still be denied of course, for example an epiphenomenalist could maintain the view that the *feels* have no functional role.
[28] Some foreshadowing is warranted here; the illusionist position is roughly that there are not some extra *feels* on top of our various behavioural reactions, judgements, attitudes, and so forth, that these types of things account for the *feels*.
[29] Or of life, heat, morality, maths….

solving the hard problem will require radical change to our scientific understanding, such as considering consciousness as an irreducible fundamental feature of the universe, along with features like gravity and energy[30].

The Meta Problems

Not everyone agrees that the Hard Problem is a hard problem, and the question of whether there even is a Hard Problem is the Meta Hard Problem of Consciousness (Clark 2014, p.272). For example, Churchland (1996) considers that the Hard Problem is built on an argument from ignorance: that those advocating for it are simply not able to know whether it will prove to be particularly challenging to solve how phenomenal consciousness is produced in our brains in comparison to other neuro-psychological challenges, let alone whether we need to depart from our physicalist worldview to solve it. As Churchland (1996) emphasises, we need more than imagining there is a metaphysical gap to have sufficient reason to believe there really is a gap. As it stands, it is not entirely clear there is enough conceptual clarity about what we really mean by 'phenomenal consciousness' (Churchland 1996; Clark 2014, p.259; Mandik 2017; Niikawa 2021, p.18; Levy 2024, p.20) to begin to know whether it is a hard or easy problem.

Another meta-problem is the Meta-Problem of Consciousness: the problem of why we think there is reason to believe that there might be something like the Hard Problem of consciousness, and why it might be hard (Chalmers 2018). For example, what leads people such as Chalmers (2018, p.7) to say things like "It is hard to see how

[30] But not life, heat, morality, maths…

consciousness would be physical". This is the problem-problem; the problem of explaining why we think there is an explanatory problem with regards to how experience is produced by the brain.

These two meta-problems are related. If there is no Hard Problem, then the reasons for why we think there is a Hard Problem—which would be answers to the Meta-Problem of Consciousness—will be merely psychological. The answers will only be informative with regards to how we think about the metaphysical issues, not anything about the metaphysical status of the issues themselves. If there really is a Hard Problem, then answers to the Meta-Problem of Consciousness will likely split between psychological explanations and metaphysical explanations. Conversely, the existence of psychological answers to the Meta Problem of Consciousness—why we think there is a problem of consciousness—could explain away and fill in the explanatory gap. This is where illusionism fits in.

3 Illusionism as a Solution to the Hard Problem of Consciousness

In responding to the Hard Problem of Consciousness, Frankish (2017) outlines three broad approaches to how phenomenal consciousness may relate to physicalism: Radical Realism; Conservative Realism; and Illusionism (which also provides an answer to the Meta Problem of Consciousness). These views broadly cover the range of possible approaches to fitting phenomenal consciousness into a scientific worldview to solving or dissolving the hard problem.

Radical Realism: Radical realists emphasise the incompatibility between what phenomenal consciousness seems to be and our established scientific worldview (Frankish 2017, p13). They point to the apparent anomalousness of the content of phenomenal consciousness, and how these aspects do not seem to have a functional role. Radical realists respond to this incompatibility by accepting the reality of phenomenal properties (*redness* and so forth) and advocate for a radical revision of the scientific worldview to accommodate them (e.g. Chalmers 1996 and Goff 2017). Substance dualism and panpsychicism (e.g. see Goff, Seager, and Allen-Hermanson 2022) are views which belong to the radical realism camp.

Conservative Realism: Conservative Realists, as their name suggests, adopt a more conservative view than radical realists in trying to reconcile phenomenal consciousness with physicalism and accepted science. When confronted with the apparent conflict between phenomenal consciousness and physicalism, conservative realists attempt to reconcile the two without making much change to either. This approach is more

scientifically conservative, as it is a better fit with current scientific understanding. Rather than needing to revise a broad swathe of science, perhaps having to discover new forces or substances, conservative realists propose that only a small part of it needs revising. This would leave us with much less work to do, compared to Radical Realism.

Illusionism: Illusionism is the view that phenomenal consciousness is an illusion caused by a systematic 'misrepresentation' of physical properties as 'phenomenal' ones (Frankish 2017, pp.13-14). Thus, under illusionism, any apparent anomalousness of phenomenal consciousness is to be explained in physical terms, drawing a distinction between the appearance and the reality of experience. Illusionism relies on this distinction between that which represents, and what is represented. According to illusionism, *how* the representation appears is an illusion.

Illusionism is motivated by the need to reconcile the features of phenomenal consciousness with what we know about how the brain and the world work, within a broadly physicalist worldview. Illusionists attempt this reconciliation by arguing that our experiences only *seem* to have the features they do, like the 'feels' such as *redness* and *painfulness*, or the supposed second-order properties. By recasting what we think the problematic features of experience *are* without changing how they *seem* to us, illusionism aims to fit the appearance of phenomenal consciousness into a physicalist worldview by saying we are wrong about our experiences: phenomenal consciousness is an illusion.

Illusionism denies that the *feels* of experience exist but acknowledges that they seem to exist (Frankish 2022, p.2). Their appearance is the result of misrepresentations or a 'distorted' perspective of our experiences (Frankish 2017 and 2022), analogous to a perceptual illusion. Perceptual illusions are stable perceptions that represent things as different to how they really are and can often only be dispelled by viewing the object from a different perspective. According to illusionism this is an apt description of conscious experience. A conscious experience may seem a certain way to us, as though it has certain properties, but the reality is quite different. Conscious experiences are like rainbows (Frankish 2022 and 2023); there *is* a reality to them (droplets in the sky refracting light), but the way they seem (a bridge of colours in the sky) is not real. More broadly, this fits with our wider concept of illusions, where perceived appearances do not correspond to the reality of their objects. Typically, illusions persist despite our awareness of them. No amount of thinking about or knowing about an illusion will make it go away, instead we need to adopt a different perceptual perspective to break the illusion. However, when it comes to phenomenal consciousness, this strategy of adopting a different perspective is not available to us, and we are limited to our viewpoint from 'the inside'[31].

To be clear, Illusionists do not deny that we have conscious experience, nor do they even claim that consciousness itself is illusory (Frankish 2022b). The illusionist position is that the 'phenomenal properties'—like *redness* and *painfulness*—are illusory (Frankish 2022b). Illusionists hold that the *feels* "are not distinct from functional and

[31] "How could they see anything but the shadows if they were never allowed to move their heads?" – Plato, *The Republic*

representational properties, are not clearly revealed to introspection, do not resist scientific description, do not present a deep explanatory problem" (Frankish 2023, p.6). Taking an illusionist view of consciousness allows us to take the appearance of experience seriously, while avoiding some of the problems associated with phenomenal realism (Ross 2016).

Some illusionists think that belief in phenomenal consciousness amongst philosophers is a kind of theoretical mistake. This mistake is brought about by believing in a self, a kind of observer, that is undergoing the experiences (Blackmore 2022, p.45), which is psychological baggage from previous ways of thinking, part of what Dennett calls the 'Cartesian Theatre' (Dennett 1993).

Another view is that the mistake of believing in phenomenal consciousness arises from a confusion between the cause and the object of what we get from introspection (Dennett 2020). The point here is that intentional objects and the causes of them are separable. If we're imagining something, say a dragon, then the cause of the imagined thing is different to the imagining. Not recognising this when we consider phenomenal things, such as the *redness* of an apple, is a root cause of the theoretical mistake. Place (1956, p.49) considered belief in phenomenal properties to be a logical mistake "of supposing that when the subject describes his experience, when he describes how things look, sound, smell, taste or feel to him, he is describing the literal properties of objects and events…".

Much of our understanding of a phenomenon depends on our perspective, both our physical perspective in terms of lines of sight and what we can see, and our theoretical

assumptions, as perhaps exemplified by this quote from Anscombe about a conversation with Wittgenstein:

> *He once greeted me with the question: "*Why do people say that it was natural to think that the sun went round the Earth rather than the Earth turned on its axis?*" I replied: "*I suppose, because it looked as if the sun went around the Earth.*" "*Well,*" he asked, "*what would it have looked like if it had looked as if the Earth turned on its axis?*"*
>
> - Elizabeth Anscombe, *An Introduction to Wittgenstein's Tractatus*

Illusionism is an attempt to overturn our existing theoretical perspective, to dissolve the Hard Problem of Phenomenal Consciousness. Drawing on Anscombe's example, we recognise that we are psychologically (and perhaps culturally) predisposed to certain explanatory views, but potentially unpalatable alternatives—illusionism, the Earth turning on its axis—are available and should be seriously considered if they better align with the empirical evidence.

3.1 The Second-Order Properties and Physicalism

Frankish (2017, p.15-16) draws a distinction between weak and strong illusionism. Weak illusionism is the view that while the *feels* of experience (*redness* and *painfulness* and so on) are real, the second-order properties (Frankish refers to them as features) are illusory, or at least do not have any problematic features with regards to physicalism. Strong illusionism encompasses weak illusionism and adds on the view that the *feels* are also illusory. Before moving onto arguments for strong illusionism, I consider the supposed issues for physicalism due to the second-order properties of conscious experiences.

The second-order properties of conscious experiences are a focus in the debate around the Hard Problem of Consciousness. However, they are generally (i) ill-defined (or, perhaps, elusively defined), (ii) not necessarily a challenge to physicalism, and (iii) could plausibly be illusory. In this section, I argue that these claimed second-order properties do not pose a strong challenge to a physicalist worldview, as they can either be discounted or reconciled with physicalism. However, since second-order properties are not definitional of phenomenal consciousness, the challenge of accounting for the claimed phenomenal properties—the *feels* of experience—will persist.

The phenomenal properties are sometimes considered to be simple, ineffable, intrinsic, private, and immediately apprehended (Frankish 2017, p13). I take 'simple' to mean that the experience of phenomenal properties cannot be subjected to further decomposition. Ineffability refers to the impossibility of expressing them in language: our apparent inability to get someone else to understand what we mean about phenomenal experiences unless referring to a relevant experience they have undergone themselves (Dennett 1988). By something being intrinsic, it is meant that it is non-relational (Dennett 1988). The privateness of phenomenal consciousness means it is not publicly accessible, making interpersonal comparisons problematic (Dennett 1988). Phenomenally conscious experiences are thought to be immediately apprehended: in some sense directly accessible, or with which we are directly or intimately acquainted (Dennett 1988).

It is not clear that the attribution of these second-order properties identified by Frankish (2017) are widely endorsed in contemporary philosophy, at least not as a

complete set. Schwitzgebel (2017) constructs what is intended to be a modest starting point for theorising about phenomenal consciousness, lacking problematic assumptions, which does not even include any of the second-order properties. Similarly, Mandik (2017) argues that there is insufficient content in the theoretical terms used to discuss phenomenal consciousness—such as qualia, phenomenal properties, and 'what it is likeness'—such that "nothing worth saying is said by employing any of [them]" ('qualia quietism').

Even if we were to assume that there are these second-order properties of phenomenal consciousness and they are close to the set that Frankish and others (e.g. Chalmers 2018, p.49) provide, it is unclear how the properties are a challenge to physicalism. In the following subsections I will consider each of the supposed properties, and whether they are necessarily a problem for physicalism.

3.1.1 Simple

Phenomenal properties of experiences are often described as simple, atomic, or primitive, implying they cannot be decomposed into simpler components. Consequently, they are considered unanalysable (Shabasson 2022, p.446). However, just because they seem simple to the introspecting subject does not necessarily pose a challenge to physicalism, so reconciliation is possible. Our direct access to the workings of our mind is limited, often confined to the outputs that reach consciousness. This limitation likely contributes to the perception of these properties as simple. We are unable to introspectively decompose or query these results, constrained as we are by the inability for us to alter our perspective, resulting in the brute appearance of

simplicity. Nonetheless, we understand that these experiences are not truly simple. For instance, we know that the brain's visual processing system, when manipulated or damaged, alters our visual experiences.

Consider the interaction between hearing and vision in speech perception: seeing people speak influences our auditory perception, yet we do not perceive this as the integration of two distinct sensory channels. This phenomenon is vividly illustrated in experiments by McGurk and MacDonald (1976), where manipulated sound and lip movements lead us to experience hearing sounds that were neither spoken nor mouthed. While our experiences appear as a unified whole, we know that they (or at least some of them) are compositional.

In addition to the non-simple nature of the perceptions that lead to the *feels*, the *feels* themselves seem to be at least somewhat analysable. We can describe the shade or vibrancy of red or the degree of pain. A particular pain in my leg can be mild, fierce, throbbing, itchy, intense, burning, or stabbing. While a case could be made that each particular instance of pain has a distinct and simple *painfulness* phenomenal property, the case could also be made that the experience of *pain* has a kind of structure to it.

Phenomenal experiences may seem simple, but empirical evidence indicates that at least some are not. Even if certain content were genuinely simple, it remains unclear why this simplicity should be regarded as metaphysically special or in conflict with physicalism, as opposed to a psychological fact. Their apparent simplicity could merely be a matter of our perspective on them or how they appear to us. If we do eventually reach a point where some aspect of a *feel* cannot be further decomposed, it is not

clear what the metaphysical challenge is. We should not expect introspection to give us the ability to endlessly analyse our experiences in ever finer levels of detail.

3.1.2 Ineffable

The *feels* of experience are considered to be ineffable. When I see a vibrant red apple, feel the ice in a bucket, or experience the sensation of hunger, I rely on shared examples to describe these phenomena to you. If you do not know what red is, telling you that the apple is "really, really red" is futile. Similarly, explaining that a nagging, unpleasant feeling deep in your belly is hunger is unhelpful if you have never experienced hunger. Without reference to food, you might mistake the sensation of hunger for a stomach-ache or anxiety. Furthermore, our experiences may not directly align with what should be the equivalent experience felt by others[32]. All this means that we are not able to meaningfully describe *feels* of experiences to each other, we can only point to shared or similar experiences.

We can describe our experiences somewhat, as with the example of the *pain* in my leg potentially being mild, fierce, throbbing, itchy, intense, burning, or stabbing. It is not entirely clear to me what further level or kind of description is being sought by those who say experiences are ineffable. However, even if my attempts fall below the called for standard of description and fail to provide a description without making use of a reference to some shared experience, moving from the inability to describe something

[32] The awe I feel looking up at the stars is an awe of their size, distance, age, and appreciation that in the context of stars we are tiny beings. This is likely a very different awe from that which is felt by people who do not know that the stars are suns which are very very far away.

to asserting that it possesses an unusual non-physical property is a significant leap. The ineffability of these experiences could stem from structural issues, such as our natural perspective or different content formats in the brain, which limit our ability to introspectively access or communicate these experiences.

Similarly, we encounter various non-mental natural things that we cannot, or used to not be able to, adequately describe in words, yet we feel no pressure to consider them to be non-physical. For instance, the behaviour of particles in quantum mechanics is notoriously difficult to intuitively grasp and articulate in ordinary language, yet these phenomena are well accounted for in mathematical form within physics. Just because we cannot express something in ordinary language, does not mean that this will always be the case, and even if it proves to be so, it is more indicative of a cognitive limitation than the presence of something non-physical.

3.1.3 Intrinsic

Some philosophers argue that the *feels* of experience are intrinsic, meaning that they are non-relational[33] in a way that means they cannot be linked to physical properties or causal roles (Jaworski 2011, p.219). This means that if we took away all the relational aspects of a phenomenal experience (if there were any to begin with), something would remain, something over and above our reactions to it, or any relationship to anything else. For it to be worthwhile to say that phenomenal experiences are intrinsic, what is leftover must be a something that can be reasonably be identified as a

[33] There are multiple notions of intrinsicality (Marshall and Weatherson, 2018), but generally the philosophical discussion about qualia seems to settle on relational versus non-relational properties.

phenomenal experience. It would not be enough to say some part of the composition of the *feel* or its construction has an intrinsic property. The *feel* itself is supposed to be intrinsic.

It is not clear, however, that there is anything intrinsic about phenomenal experiences or why we should suppose they are intrinsic. The claim that phenomenal experiences are intrinsic seems to arise from intuitions, and I do not believe there is empirical support for it. We should doubt that our experiences are non-relational because we know that many of them are dependent on sensory input—or at least in the case of imaginings and hallucinations the output from brain processes that normally deal with sensory input—and their relationship to contextual and conceptual content. For example, feelings of anxiety can be reappraised as excitement (Books 2014), and feelings of bodily warmth can be comforting or uncomfortable depending on the ambient temperature. These examples suggest that the *feels* are relational, but even so, whether there is anything to them over and above the physical properties or causal roles is a primary point of contention with regards to illusionism and is covered in more detail later in this paper.

3.1.4 Private

Phenomenal properties are thought to be radically private (Frankish 2017; Niikawa 2021), in that only the possessor of the property can know or experience it. Nobody can 'see' inside our minds to experience what we are experiencing. However, this privacy may be epistemic rather than metaphysical. Having private mental events is like each of us having a box that only we can see into (following Wittgenstein 2009 (1953),

p.106). You do not know whether I have a beetle or a barnacle in my box. Because you cannot look inside my box you must rely on my reports about the contents, and my observable behaviour. There is no mystery here, only that you are unable to obtain and physically open and look inside my box. At some point, that may change. Similarly, in the future, neuroscientists may be able to share experiences between brains. If conscious experiences are brought about by brain states, then it should be possible to read a brain state and write it to another brain, so that both have the same experience[34]. Thus, from the physicalist viewpoint, the property of privacy arises due to factors such as structure, access, and perspective. It is a property of our psychology and physiology[35], but not a special metaphysical property. As such, it is not in conflict with physicalism.

3.1.5 Immediate

Conscious experiences are considered to be immediate. The motivation for thinking of an experience as being immediately apprehended, or that we are directly acquainted with them, is understandable. Most of the time, it certainly seems that way. This should not be surprising, given our conscious experience is of a near constant feed of sensory impressions. We can even manipulate these impressions, by turning our heads

[34] It may be that in all but trivial cases the experience will never be exactly the same, but in the same way we can come to practically but perhaps not wholly share an understanding of the meaning of a word, it could be close enough.

[35] If (a big if) there were aliens whose brain-equivalents were transparently observable to their conspecifics such that patterns of activity could be externally detected and reliably correlated with phenomenal experiences, then their phenomenal consciousness would not even be epistemically private. A long history of association between observed brain states and reported *feels*, would (I believe) make it unlikely that the aliens would subscribe to metaphysical privateness or the Hard Problem.

or by tasting a different dish, with no noticeable lag between willing it so and for there to be a change to our perceptions.

Except, while it seems this way, it is not how it is. We have no sense of what comes *before* our conscious experience, how it is brought about in the brain, or what it is composed of, nor of any mediation or presentation layer. We just get finished product, and we get it without noticeable gaps or omissions, as though it were directly present. Despite how it seems our conscious experience is not immediate. It has delays.

For instance, sometimes when we try to recall something or try to figure out why we came into the kitchen, it takes some time to come to us. Or we will look at an image, and it will take some time for it to 'resolve' and be recognisable. Consider the case of looking at an image, and it taking a while to 'pop' and become clear what it is of—I sometimes have this experience with maps when it is not clear which areas are meant to be sea and which are land—we have the complete visual experience of the map image in terms of what there is to see, but it is not until some top-down process recognises and marshals the relevant concepts, that we experience the visual phenomena as a recognisable map. These are cases of delays caused by cognitive processing.

Other examples tell against the idea that the *feels* of experience are immediate. Our visual system smooths out the gaps in our vision when our eyes move (Irwin *et al*. 1998), but we do not notice the gaps. We are aware that the speed of sound is slower than the speed of light, such that we see a distant event before we hear it. However, visual stimuli are processed in the brain more slowly than audible stimuli (Shelton &

Kumar 2010; Jain *et al.* 2015). This means that if visual and auditory stimuli reach our brain at the same time, we will be conscious of the sound before the image. This leads to there being a sensory cross-over distance. Less than that distance and we will hear events slightly before we see them, because the auditory information is processed faster. Beyond the cross-over point we will see events before we hear them, because the light from the event will reach us much more swiftly than the sound and the amount of time it takes to be processed. Consequently, our conscious experiences are neither immediate nor do they always align temporarily with the events they represent.

Some might object that my response concerns the production of conscious experience, not the apprehension of it. However, conscious experience *seems* immediate because that is all there is for us. All that comes before is dark and impenetrable, unknowable to introspection. It is not clear that our phenomenal consciousness could be experienced in a way other than one that seems immediate. Consider how we do not experience the gaps in our vision during eye saccades (Zimmerman and Lange, 2022), and how we do not experience the blind spot in our visual field. We experience what is made available to us to experience, and just as we may lose track when driving or asleep, we do not have an awareness of those periods when we are not conscious. We do not experience the gaps, so our occurring experiences always seem immediate. This immediacy property, therefore, may not have any significant metaphysical implications, being 'merely' a feature of the way our mind work.

3.1.6 Summary

While experiences may seem to possess these second-order properties—e.g. seeming to be ineffable, immediate, and so on—this could be an illusion. The way the brain functions could give rise to an illusion of these second-order properties, leading to 'problem intuitions' (Chalmers 2018). For instance, this illusion could arise from the modularity of the brain; we only get the experience as an output, we have no access to how it came about or what it is composed of. Alternatively, I may naturally (mis)remember my experiences as having ineffable phenomenal properties despite that not being the case.

Given all the above, it is not clear that the answer to the Meta Hard Problem (Clark 2014 p.272)—whether there is a Hard Problem—is 'yes'. While intuitively compelling, and a difficult explanatory gap to bridge, there is reason to doubt the Hard Problem on the basis that we should be optimistic about the progress of science and our intuitions may not be a good guide to what science can ultimately explain (Churchland 1996). It is not clear why the supposed second-order properties should be a threat to physicalism or suggest that phenomenal consciousness is separable from functional goings on in the brain. However, while there seems to be plenty of reason to doubt that the Hard Problem is a real problem, it still needs to be taken and responded to seriously, particularly when it comes to the phenomenal properties, such as *redness* and *painfulness*. Next, I turn to arguments for strong illusionism which are directed at the nature of phenomenal properties.

3.2 Arguments for Illusionism

The motivation for non-radical physicalists to accept illusionism seems clear: if phenomenal consciousness does not fit into our physical worldview, then perhaps we are mistaken about it. But why accept illusionism? Here I outline two positive arguments Frankish makes for illusionism in answer to this question. These arguments are: “Phenomenal Consciousness does not need Explaining in Non-Physical Terms”, and “The Argument Against Anomalousness”.

3.2.1 Phenomenal Properties Do Not Need Explaining in Non-Physicalist Terms

Frankish’s first argument (2017, p.27) is that we do not need to explain the *feels* of experience—the phenomenal properties—in non-physical terms:

1. (i) If people have beliefs about X and (ii) these beliefs can be fully explained without requiring the existence of X then (iii) we can discount X as an illusion
2. (i) Some people have beliefs that our conscious experience includes phenomenal properties
3. (ii) These beliefs can be fully explained by mental processes, and do not require the existence of phenomenal properties
4. (iii) Therefore, we can discount beliefs in phenomenal properties as a mistake (i.e. it is an illusion)

Unstated in Frankish’s argument is that if an X seems problematic in how it fits within a physicalist worldview, then we have additional reason to seek to reject or find

alternative explanations for it. The application of the argument is comparable to our approach when encountering a magic trick. We do not need to explain 'the magic', because it would be an extravagance in terms of upturning commonly accepted science, but also because we can explain the magic away in other terms, such as it being a clever trick or illusion.

The idea here is that we should treat phenomenal properties as we would treat an illusion conjured by a magician. We can explain away the phenomenal illusion because our various beliefs and behaviours (e.g. verbal reports) about phenomenal properties can be accounted for by physical mental processes (Dennett 1979, p.95 and Dennett 1991, pp.363-364). Even amongst non-physicalists, this is generally held to be the case (Frankish 2017, p.27). Though I do not want to raise philosophical zombies as part of the general argument I am making, it is worth noting that non-physicalists who find philosophical zombies conceivable may be at least implicitly endorsing the position that our beliefs about phenomenal properties can be fully explained by mental processes, even in the absence of the phenomenal properties. Frankish's argument aims at removing unnecessary entities. Its general force is that we can discount things—like gods or ghosts—if we can explain everything attributed to them—like rainbows, or moaning noises coming from the direction of pipes—via other means.

Phenomenal Consciousness is a Datum

A response that can be made to this argument is that it does not hold because the phenomenal properties of experiences (*redness* and *painfulness* and so forth) are a thing to be explained, rather than an explanation, so our beliefs about them do not

need explaining (Chalmers 1996, p.187-189). The idea here is that phenomenal consciousness is a datum, an axiomic starting point that we should all accept as true. Just because we can explain away our beliefs about phenomenal consciousness does not mean it can be discounted like ghosts or gods, as we have independent reasons to believe in its existence. Chalmers' point is that phenomenal consciousness (and in particular the phenomenal properties) does not need to fit into an explanatory framework like a law of physics for us to accept its existence, it just is. It is an observable feature of the world, rather than a theoretical posit, therefore it cannot be explained away. Even if we can fully explain everything else without needing to make use of phenomenal consciousness, instead of being banished it remains as something that still needs explaining. Based on this view, we cannot discount our beliefs that our subjective experience has phenomenal properties as a mistake (as per #4 above), because our experiences do include phenomenal properties like *redness* and *painfulness*. Analogously, while we can discount gods or other supernatural entities as the cause of rainbows, we cannot discount rainbows; they're a thing to be explained, so cannot be explained away.

This kind of response would hold if it were the case that our judgements or beliefs about phenomenal consciousness did not fully exhaust the experience of it, if there was something further to explain. Frankish's response to Chalmers (Frankish, 2017, p.30) is to point out that illusionists deny that phenomenal properties are real, but they accept the introspective awareness that we seem to have them. This is akin to accepting the appearance of a rainbow but rejecting it as a really being a multi-coloured arc spatially located in the sky (Frankish 2022 and 2023). Although we can

reliably cause their apparent physical manifestation, say with a hosepipe spraying water on a sunny day, there is no physical thing that corresponds to the rainbow. A rainbow has an apparent virtual location, but not an actual location, because rainbows are not really there. We experience rainbows because of the way the different wavelengths of refracted light are processed and represented by the brain. We can accept the appearance of rainbows, but not their reality *as they seem to us*. The illusionist move does not deny that we have a kind of experience that demands explanation, rather it changes the nature of the thing to be explained and therefore the range of suitable explanations. We are left with explaining why it *seems* there are phenomenal properties, instead of explaining how there could be phenomenal properties and how they come about. These seemingly phenomenal—'quasi-phenomenal'—properties can (presumably) eventually be explained in terms of mental processes.

The Phenomenal Realist Response

The phenomenal realist could respond to Frankish by claiming we are in some way directly acquainted with our phenomenal experiences and their properties (Frankish, 2017, p.30), rather than being aware of them via interpretation or inference. This direct acquaintance would give us a secure epistemic access to our phenomenal properties, which warrants us to say they really exist.

However, this response cannot rely on how phenomenal consciousness appears to us subjectively to have any force. The illusionist can reply that the appearance of direct acquaintance is simply that, an appearance. Everything to do with how phenomenal

consciousness seems to us comes to us via their appearance, how they are represented to us. As with a visual illusion, to say what we perceive is not an illusion, and to say we can tell it is an illusion because we're directly acquainted with it is not sufficient and additional reasons are required. To be successful, the phenomenal realist's response needs to make a case on objective grounds that we have special access to these properties, a kind of access that is resistant or immune to the illusionist claim. Illusionism has the advantage over realism, being both physically and psychologically conservative.

Summary

Broadly then, Frankish's argument is that phenomenal properties do not need explaining, because we can in principle explain how mental processes make them *seem* to exist. They are quasi-phenomenal rather than actual phenomenal properties. The hard work of explaining how these quasi-phenomenal properties occur remains ahead. This argument can be supported by the broader case that Frankish makes that we should be motivated to consider illusionism, because—unlike classical phenomenal consciousness—it fits within our scientific understanding of the brain and the world.

3.2.2 The Argument Against Anomalousness

Frankish's second argument (2017, pp.27-28) is that we should accept illusionism as a simpler, and therefore more compelling, explanation than the alternatives:

1. If an [observable] property or feature resists explanation in physical terms or is detectable from only a certain perspective, then as a matter of parsimony (and

experience) we should consider it to be illusory, rather than being an anomaly in our physical worldview

2. Phenomenal properties resist explanation in physical terms
3. We can only detect phenomenal properties from introspection; we can't detect them from other perspectives, *and* they resist explanation in physical terms
4. Therefore, we should consider phenomenal properties to be illusory

The first premise is key in this argument, with two important disjuncts to consider; resisting physical explanation, and only being detectable from a certain perspective.

Resists Explanation in Physical Terms

The first disjunct in premise one presupposes physicalism. Frankish's main case for physicalism, or rather his case against radical realism, is a) he presumes that conservatism is preferable in science, and b) this should be especially the case when the case for radicalism is based on a limited and distorted perspective (Frankish, 2017, p24), as we should consider our own experience of consciousness to be.

Just because an observed property resists physical explanation does not mean that it is illusory, or even that illusionism is the simplest explanation. For centuries astronomers struggled to come up with a good explanation of the motion of the stars in the sky (or even what they were) until the development of the heliocentric model of the solar system. Before this development it would have been premature to declare that the stars or their motion were illusory. Similarly, at a point in time when our physics did not include radio waves, if one encountered a 'talking box'—a radio—it would (initially) resist explanation in physical terms. To dismiss Radio 4's *Thought for the Day* as an

illusion may be tempting but would be a mistake[36]. What Frankish means here in his argument is not simply that something is challenging to explain in physical terms, but when something seems like it will not be described in physical terms, i.e. we are inclined to think that there is a 'Hard Problem'. Frankish points to a track record in science and philosophy of being able to explain in physicalist terms the previously unexplainable. We now know why celestial bodies move, or appear to move, as they do from our perspective, and we can explain things like how (if not why) we are listening to *Thought for the Day*. This track record of explaining the previously unexplainable, and the general banishment of non-physical terms (such as gods and ghosts) in our explanations gives us good reason to be confident that if something observable seems non-physical, then that probably is not the case, and it is more likely that it is an illusion.

Only Detectable from a Certain Perspective

The second disjunct of interest is the main illusionist move. We are familiar with a range of sensory illusions whereby things seem extraordinary—such as stage magic or visual illusions—but prove to be a kind of trick exploiting our limited perspective, or to have been brought about by a quirk of our brain processing. These extraordinary seemings can vanish when seen from a different perspective. This is the case with many magic tricks, or if we break up the view of a visual illusion into smaller pieces. Animations, cinema, and flipbooks, for example, all exploit the Phi Phenomenon to induce a sense of motion, while being constructed out of a series of still images. When

[36] This is reminiscent of Arthur C. Clarke's third law: "Any sufficiently advanced technology is indistinguishable from magic".

we slow the sequence of images down, we can see the component parts of the illusion, and how a different temporal perspective gives rise to an illusion of motion. Or consider Figure 1, below, which shows a peripheral motion illusion. The design of the image exploits our perceptual system to make parts of the image appear to be rotating when not being directly looked at. However, there is no actual motion, just the appearance of motion, and if one shifts perspective to look directly at portions of the image the appearance of motion disappears. Overwhelmingly, things that seem extraordinary turn out to be mundane (or at least still rooted in physicalism) when we investigate them from a different perspective.

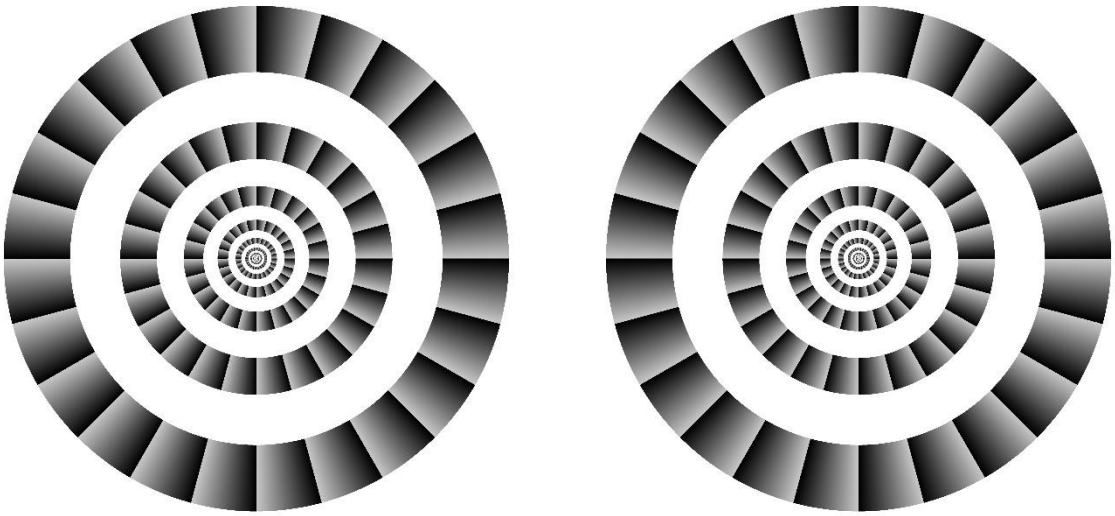

Figure 1 – A Fraser-Wilcox peripheral drift illusion (Source: Kitaoka 2012)

Via inductive argument based on our general experience of anomalous things, or endorsement that mental phenomena are grounded in physical stuff, we should accept that the apparently anomalous *feels* of experience will ultimately also be understood in the same way. Anomalies, like magic tricks and visual illusions, only seem to be beyond

physicalist explanation until we understand what is going on. What we have with the *feels* of experience is something that seen from our limited perspective, from the inside as it were, seems to some to be rather extraordinary in the non-physical sense. When we see what seems to be the causal machinery of experience in action, we should accept that the non-physical extraordinariness collapses into a physicalist explanation.

Illusions

An illusion is something that is *systematically* "seen as it is not" (Fish 2010, p.3) or appears falsely (Mandik 2017, p.145) from a given perspective, but is present with a different appearance when viewed from a different perspective. An illusion occurs when a set of features observed from a certain perspective appear substantially different in a way not congruent with the component features. The stage magician's show fails to impress when one can see all the contrivances—or they are not angled just so—to make the magic work. Visual illusions that make static images appear to move can be dispelled by partially covering them up or angling them differently. What Frankish is arguing for is that while it may appear that there are phenomenal properties, they do not actually exist, their apparent existence is an illusion based on our perspective on how experience is being (mis)represented.

We have now circled back round to the supposed explanatory gap. We can observe things in the brain that we have good reason to believe play a causal role in bringing about experience and the appearance of phenomenal properties like *redness* and *painfulness*. To some, these things, these patterns of activity being realised by neurons,

are too far removed, too different from our conscious experience, to be the same thing from a different perspective. There is, in short, an explanatory gap. To others the gap is there but bridgeable. To the illusionist, the structures of neurons firing in the brain are the publicly observable view, and our inner-introspective experience of them gives another view, one that is based on useful but not accurate (mis)representations.

Humphrey (2012) suggests the illusion of consciousness is possible due to its perspectival nature. He introduces the Gregundrum (Humphrey 2012, pp.5-7), which is a real object which when viewed from a very particular perspective looks like a shape that is physically impossible (see the figure below). Being constrained to a single perspective on our experiences allows for there to be just such an illusion to be taking place. This limited perspective could give rise to the *feels* of experience and the various putative second-order properties. For example, our conscious experiences may seem simple because we are presented with whole outputs with no access to the workings that produced them.

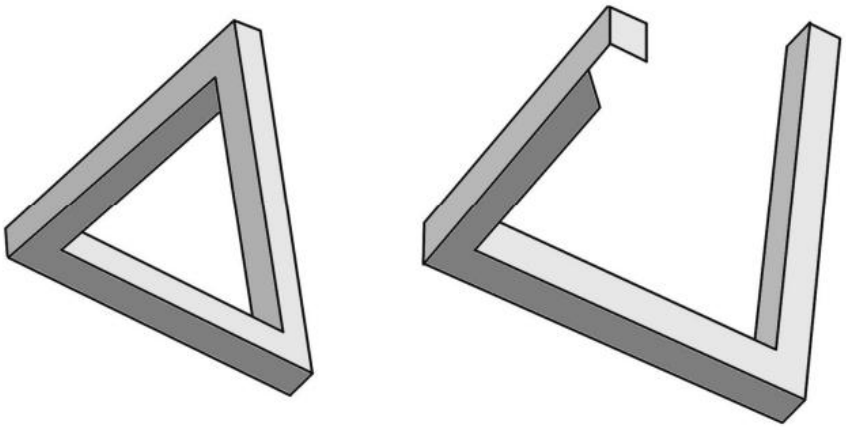

Figure 2 - Left: the impossible Penrose triangle. Right: the possible Gregundrum which can appear as a Penrose triangle from the right perspective (source: Frankish 2017)

Our perspective on conscious experience arises from our viewpoint ‘from the inside’. Experience is ‘presented’ to us for action. We get what might not be ‘finished product’—many of us have gazed at things like maps before not understanding what we are seeing until **pop** something in our brain works it out and we find ourselves perceiving what is obviously a map—but we get an output that has been processed by activity in the brain. We do not get to see early raw inputs, we do not get to see inside the box to see the workings, how the magic is done, we just get some outputs. We cannot ‘from the inside’ view those outputs differently, they are cognitively opaque. We can only provoke new outputs, by changing the outside world (perhaps rotating an image to see it in a different way) or changing our perspective. But though we can change what we will experience, we cannot change our perspective on our subjective experiences.

While we are not observers in a Cartesian Theatre, as evolved beings we should expect that the way we experience the world, and the outputs from the workings of our brain, should be *useful* but not *accurate* about the nature of things. What supports their usefulness, are those salient features or variations of the objects that are relevant to our behaviour and survival and how we track them over time. Useful appearances do not need to match reality. This idea is the general thrust of my main argument in support of Illusionism and will be expanded upon in more detail in the next section.

4 An Argument for Illusionism: We Should Expect the Nature of Conscious Experience to be Useful Rather than Accurate

> *Brains are survival engines, not truth detectors. If self-deception promotes fitness, the brain lies. Stops noticing—irrelevant things. Truth never matters. Only fitness. By now you don't experience the world as it exists at all. You experience a simulation built from assumptions. Shortcuts. Lies.*
>
> - Peter Watts, *Blindsight*

In this section, I provide an argument in favour of illusionism. The goal is to lessen the attraction of the Hard Problem and to make illusionism seem a more plausible position. My argument is that as evolved creatures, we should expect how we perceive the world, including ourselves, to tend towards being useful rather than accurate *per se*. This extends to how our experience seems to us. As the quote above from Peter Watts claims, brains support survival and are only truth-sensitive to the extent to which it aids survival.

We should find illusionism agreeable because, beyond the appearance of subjective experience we do not have good reason to accept its representation as it appears. From introspection, we only get appearance, and we have reason to believe that the appearance is useful which may at times be at the expense of accuracy. As Frankish puts it:

> “We have abundant evidence of the unreliability of introspection, and there is no reason why an evolved cognitive system should represent its internal states to itself in a transparent way, as opposed to an adaptively useful one” (Frankish 2017b, p.279).

First, as evolved beings, we should expect the appearance of our subjective experience to be like our various other capacities and features: adapted and fit enough for its purpose, but not necessarily 'perfect' or 'ideal.' Just as with the limitations of the eye, including those which give rise to the visual blind spot, our limited lifespan, the risk to our lives from growing wisdom teeth or from giving birth, the architectural follies like the wandering backtracking vagus nerve, and the vestigial appendix, so with the mind. We should expect the workings of our minds to (hopefully) be ‘good enough’ for what works for us over evolutionary timescales and niches, but we should also expect shortcomings and issues with those workings.

Secondly, this is generally what we find: how things appear to us is sometimes not accurate, yet we seem reasonably fit and well-adapted to interacting with the world despite these inaccuracies. For instance, our conscious experiences may seem simple because we are presented with whole outputs with no access to the workings that produced them, but like someone at the top of a large organisation, perhaps the output we need is the ‘executive summary’—the outcome of the processing—rather than the low-level details

Is the appearance of our subjective experience, the *feels*, useful despite the inaccuracy? Or is it useful because of the inaccuracy? There will be cases where the answer is 'despite the inaccuracy,' and others where it is 'because of.'

The 'useful rather than accurate' argument I am making is a counter to an argument which motivates the Hard Problem (constructed from Chalmers 1995)[37], which is one of the chief targets of illusionism:

1. If physicalism is true, then everything can be fully explained by physical facts
2. Subjective conscious experience is not (or won't be) fully explained by the physical facts (as we know them)
3. Conclusion: physicalism (at least as currently understood) is false

I will be taking aim at premise #2 of the argument above. What we think about our conscious experiences is in large part based on their appearance: how they seem to us. We are reliant on this appearance as a basis to theorise about the nature of our experiences, even though those appearances may mislead us as to their nature. This is a somewhat soft ground—and a limited perspective—from which to judge the reality and nature of our experiences. We have plenty of reason to consider our introspections and experience to not always be a good guide to how things really are, let alone be a reliable guide to whether the appearance of experience is fully explainable by the physical facts or not. If nothing else, given our experiences involve perceptions, the

[37] Note that in this formulation there is no outright rejection of physicalism, as Chalmers is a property dualist rather than a substance dualist. Under this view, conscious experience will in some way be dependent on or caused by 'physical stuff', but this will involve a distinctly different kind of property—mental properties—rather than just physical properties.

existence of perceptual illusions should make us cautious as to whether we are conflating appearances and reality when it comes to subjective experience. Here is the outline of the argument I will make:

1. We should not rely on a faulty means to determine the true nature of our experiences
2. For this purpose, our introspection is such a faulty means; while it tracks features that are useful to us, it is not a reliable guide to the true nature of our experiences – it is useful rather than accurate
3. Conclusion: We should not rely on introspection to determine the true nature of our experiences, and specifically not about feels like *redness* and *painfulness*

The view is that the appearance of our experience is useful insofar as it leads to useful predictions and interactions with the external world, particularly relating to the kinds of features present and relevant during our evolution. However, it is not an accurate guide to what those features are, or what they are really like, and this extends to any insight we have about our subjective experiences. They appear as they do *to be useful*, not to provide us with insight as to the nature of things. This contrasts with views such as those held by Strawson (2018) who holds (drawing on Bertrand Russell) that "we know something fundamental about the essential nature of conscious experience just in having it".

We know from how our perceptions mislead us that we should not always rely on how things seem to us. We come to realise how misleading our perceptions can be when we learn things such as: objects are made up of atoms; rainbows as multi-coloured arcs in

the sky do not have an independent reality; that the ground we stand on orbits the sun rather than the other way around; and that the stars are distant suns. Added to this, we know people think about and seem to see and hear things that are not real.

Turning inward, we come to learn things about ourselves—about our bodies and our minds—that show us we are not how we seem, such as the existence of the unconscious mind, how we confabulate (see the first paper in this thesis), and how biological processes from digestion to cognition are not bounded by our bodies (see the third paper in this thesis). These discoveries conflict with our naïve views of the world and ourselves, which are formed by our perspective and how things appear to us.

Given that relying solely on our limited perspective has often misled us about the true nature of things—from visual illusions to hallucinations—we should not be surprised by the proposition that our experiences are not accurate guides to reality. Since it is unwise to rely on a means that is faulty, misleading, or unsuitable for determining the true nature of things, we should be cautious about accepting what introspection tells us about the nature of our experiences.

4.1 We Should Not Rely on Faulty Means to Determine the Nature of our Experiences

It is rational to (where possible) try to be more right about things, and this requires avoiding using or relying on a faulty means to determine what things are really like. Perhaps some means or other is only partly faulty, and it provides us with some use, but we should move away from it when we acquire a better means. When we have a

reliable means to determine what things are really like, we should and do (eventually) accept the appearance-reality gaps that we discover.

By describing introspection as a 'faulty' means for understanding the reality of our experiences, and the nature or realism of phenomenal properties, I do not intend to disparage it. By 'faulty,' I mean that for certain purposes, it can be defective, misleading, or limited in uncovering the true nature of things. While generally introspection is rather useful, it is not a suitable means to determine the nature of our experiences.

4.1.1 Trying to be More Right

The first premise of the argument is an epistemological normative claim: we should not rely on faulty means to determine what things are like when we have better options available to us. I take this to be a basic claim about rationality. We want to be right about what things are like, to better understand, model, or predict their behaviour.

If our means of determining what something is like misleads us or helps us to get the right answer at a rate worse than chance, then we should not use it if we have other options available to us, even if our alternative is a coin toss. It would not be rational to knowingly opt to use something known to be unhelpful, as we would be choosing to be misled. Similarly, if we have a means that we have good reason to believe is better than another, providing it is also better than chance, then we should choose to use it. Other factors may apply of course, such as the cost and effort involved in getting the right (or a better) answer. Not everyone can build a supercollider or access sophisticated

scanning equipment to assist them in getting the right answer about some matter, but generally we should opt for the better means of determining what things are really like. We want to be more right, or at least less wrong.

This applies to accepting appearance-reality gaps, something we generally do. When we discover an appearance-reality gap, we might not fully shake off the appearance and its influence on us. However, we should and generally do accept the underlying reality anyway, because we want to be more right. We accept the appearance-reality gap, even when the appearance seems in conflict with the revealed reality.

4.1.2 Accepting Appearance-Reality Gaps

When we discover an appearance-reality gap, we come to accept it. It may be the case that we continue embracing the appearance, while acknowledging that the reality is different. Or the new understanding of the reality may change how we think about the appearance, or lead to us rejecting the appearance entirely. But we tend not to reject the discovered reality, because we want to be more right.

For example, we know that heat is molecular motion, yet this explanation is psychologically unconvincing[38]. It does not account for how things seem, such as the subjective feeling of warmth, or the comfort of a warm blanket. Few people, if any, go around thinking that when a loved one gives them a warm cuddle it is all just molecules moving around. We have not eliminated our everyday discourse about warmth to reduce it to physical terms: instead, we accept both the appearance and the

[38] At least it is to the author.

reality, despite the substantial gap between the two. We may never shake off the hold the appearance has on our ways of thinking, and we may never want to even if we could, but in accepting both the appearance and the reality, we're also accepting—living with—the difficulty in reconciling the appearance with the underlying reality. After all, this is just as we do with the difference between how we view tables as solid objects and their reality as being largely empty space (Eddington 1927), between what Sellars called the manifest and scientific image (Sellars 1962).

Similarly, there are gaps in how we perceive colour. Generally, we know that colour arises not because objects are 'painted in colour', but because of how they reflect different wavelengths of light. However, psychologically, we continue to think of objects as being bearers of colour. Another gap between appearance and reality is our perceptual experience of colour constancy. Our experience of colour is derived from wavelengths of light reflected off objects, but as the light changes throughout the day, so does what is reflected. Yet, our brain adjusts our colour perception so that we perceive the same colour despite the variation in sensory input (Choudhury 2015). This is a kind of everyday illusion where how colours appear to us does not directly match up with the reality of what reaches our eyes. We accept these kinds of reality, rather than reject them in favour of the appearances, but still hold to the appearances in everyday discourse.

When we are kicked in the shin, it *feels* as though there is a throbbing *painfulness* in our leg. Yet, there is not. It is only an appearance. Our leg may have tissue damage, and nerves might be busy sending 'pain signals', but wherever the *painfulness* is, it is not in

our leg. This is made apparent in the case of phantom limb pains. For some people with a missing limb, it can feel as though the missing limb is experiencing pain. Yet, the *painfulness* cannot be located in the missing limb, because it does not exist. Sensations of *painfulness* as located around the body is an appearance-reality gap which we must acknowledge by embracing both the appearance and the reality. We can accept that the *painfulness* we experience is not really in our leg yet knowing that is neither comforting or particularly helpful when it *feels* as though our leg *hurts*.

Through investigating the world, we have discovered various ways in which it is not like it initially seemed to be. We continue to develop our ability to improve our understanding of the world, via endeavours such as philosophy and science. Experiments are set up—particularly in investigations relating to human psychology—to control confounding variables and to remove human bias from our approach to determining what things are like. They are an attempt to become more objective and more rational in how we learn about the world, to discover the reality behind the appearances. Investigation of the world via the scientific method, and well-grounded philosophical theorising, has led to the discovery of a wide range of ways in which the world is not like it seems to us. It turned out that the Earth is round, and it circles the sun. The stars in the sky are other suns. There is a hole in the middle of our visual field. We have discovered these things, and though they conflict with the way things seem to us based on perceptual evidence or our ways of thinking, we have embraced them as how things really are.

Responding to Gaps

Appearance-reality gaps arise due to some mix of our psychology, culture, and how our sensory systems work[39]. They can occur in cases where relevant information is available to us, and in cases where it is not available. The resulting gap may be one of exclusion, addition, or (mis)representation.

In some cases, we had sufficient information about how things are to have understood the reality of the situation, but even so failed to grasp or endorse it. Such is Wittgenstein's example of the Sun; how the Sun's track across the sky appears to us would be the same whether it orbited the Earth, or the Earth orbited the Sun. There is something about human psychology (or perhaps culture) that led to the default belief that the Sun goes around the Earth. We could have believed otherwise.

In other cases, we did not have ready access to the information that would allow us to appreciate the appearance-reality distinction. The Morning and Evening stars were thought to be separate celestial bodies, and it turned out this was wrong, and that they are different appearances of the planet Venus.

The visual blind spot is an example of an appearance-reality gap involving absence or exclusion, though of an odd sort. It is a lack of an appearance of a lack. We are *blind* to the blind spot. It could have been otherwise. We could have had a blank spot in our visual field, or a kind of ongoing awareness that it exists, rather than it being a

[39] Though I think it unlikely that culture plays much of a direct role in the experience of things like *redness* and *painfulness*, it is worth considering as part of the wider point about how it shapes our perceptions and concepts, which can impact on how things appear to us. Culture does, for example, have an impact on our perception of some illusions, such as people from different cultures being more or less susceptible to the Müller-Lyer illusion (Phillips 2011, pp.161-165).

phenomenon we had to discover. Instead, we have an odd sort of illusion whereby the blind spot gets filled in, as though it was not there. This is a kind of positive illusion whereby brain processes are adding in something which is not present. Another example of appearance-reality gaps created by our brains is our over-detection of faces (known as face pareidolia (Alais *et al*. 2021, pp.1-2) whether in the clouds, on toys, or (for some) on toast. There are various ways in which the way things seem to us do not reflect the underlying reality.

In all these cases however, we accept the appearance-reality gap when we have good reason to believe there is one, even if at some level its existence seems wrong and counter-intuitive to us. Most of us cannot really think of a table other than as having a solid continuous surface, partly because that is how our senses perceive it, but partly because it is useful to think of it in that way. Habitually thinking of a table being made up of lonely atoms in the void is not really an option for us. We accept that it is the case, even if we largely ignore it and go about our business in relation to the appearance. In the case of tables at least, it will nearly always be useful and pragmatic to consider them as having a continuous solid surface, even though we may revise what we think it is for something to have a solid surface.

Once we have discovered an appearance-reality gap, we do not continue to insist the way things seem to us is the way they really are. At least, that is the end state we tend towards, though it may take time for the implications of these discoveries to be fully accepted, as evidenced by examples such as a reluctance by some to accept our evolutionary origins, or that some people still believe in a flat earth. But generally, we

see an acceptance that there is an appearance-reality gap. Even though things really seem a certain way to us, we accept that this is only a seeming, and the reality is different.

4.2 Useful Rather than Accurate

Here is the case that I will be making: we should expect our brains to generally produce useful rather than accurate representations and presentations of how things are. Sometimes (perhaps most of the time) the two will coincide, as being accurate about things can be useful. However, we should expect evolutionary forces working on existing capacities and structures in the brain to result in capacities for tracking features usefully relevant to interaction with the world, rather than what is true. This may tend to make for a faulty means to determine what things are really like. At least in part this is because 'useful' is something relevant and tractable for evolutionary processes, whereas truthfulness or accuracy about the way things really are, independent of usefulness, are not. Evolutionary forces will respond to those factors which give some kind of survival or reproductive advantage, not to how things really are. Consider how our brains employ heuristic reasoning: useful shortcuts that model generally good-enough strategies, rather than more closely approximate the truth (Tversky and Kahneman 1974 and Kahneman 2012). Similarly, consider animals that mimic the warning signals of a harmful species (like predators or poisonous creatures) to deter predators, who are responding to useful signatures in the environment rather than being able to detect the underlying truth.

Given our nature and our origins, usefulness rather than accuracy *per se* is what we should expect. And that is what we generally find to be the case; our perceptions and subjective experience are useful but have various shortcomings, biases, and tendencies to generate illusions.

4.2.1 We Should Expect to Have a Useful Rather Than Accurate Subjective Experience

Generally, we should expect our various mental capacities to be useful rather than accurate. This expectation extends to our introspective awareness of our own subjective experiences, for the same reasons we expect it of other mental capacities. This is also because of the relationship between some of those capacities and our introspective awareness and its targets. The experience of *redness*, for example, is somewhat based on—or at least caused by—our sensory and perceptual capacities and the way they function.

We should expect usefulness more than accuracy because of our nature as evolved beings, adapted for survival rather than truth. In evolutionary terms, organisms develop traits that enhance their fitness—their ability to survive and reproduce in their environment. This often means that what is most beneficial is not always the most accurate representation of reality. Evolution will only track truthfulness or accuracy insofar as it is also tracking usefulness.

Another consideration of why we should expect usefulness rather than accuracy is our lack of general access to the workings of our brain, how any thought, ability,

perception, belief, etc. actually comes about. We are blind to the workings of our wider brain, only aware of what comes to consciousness, not what comes before. This limited perspective on what is going on gives us a distorted view of our own minds. What we get is processed output, something like a useful summary to support our interaction with the world, picking out useful patterns. And this is what we should expect: something to help us to interact with the world efficiently and effectively, not something that is a reliable guide to what things are really like.

Evolved Beings – Survival Engines not Truth Machines

As evolved beings, the 'purpose' or function of our various capacities is to enhance our chances of survival and reproduction (whether at the level of an individual, gene, etc.), and this will not always align with generating true or accurate beliefs, impressions, or judgements. As with other capacities, our subjective experience and its character would have arisen from evolutionary processes of blind trial and error, to enhance our interaction outcomes with the environment, other agents, and perhaps our selves[40]. This does not necessarily require an accurate representation, or for there to be actual phenomenal red to produce an appearance or seeming of *redness*: "…a fancier style of representing is advantageous *so long as it is geared to the organism's way of life and enhances the organism's chances of survival.* Truth, whatever that is, definitely takes the hindmost" (Churchland 1987, p.549).

[40] There are some who consider consciousness to be an evolutionary spandrel, e.g. see Robinson, Maley, and Piccinini (2015).

Evolution, in terms of heritable advantageous traits that improve the chances of successfully having (successful) offspring, can gain traction on usefulness, but it cannot do so on truthfulness except insofar as truthfulness and usefulness align. They may well frequently align, but not always. Evolutionary processes will respond to features and events that have an impact on survival and reproduction; the true nature of something will not fit into this category. Rather the kinds of things that will play a role are those like recognising whether something is likely to be a threat, whether something could be a useful tool, whether something might be good to eat, or being able to dodge an incoming projectile.

To expand on the point, in the case of dodging an incoming projectile, accuracy in perception does not necessarily mean an accurate understanding of the underlying physics. For survival purposes, it is sufficient for our brains to respond quickly and appropriately to relevant stimuli. This response relies on fast, heuristic processing rather than precise calculation of trajectory, speed, and distance. Thus, while the reaction appears accurate in terms of outcome (successfully dodging), it is based on practical usefulness rather than a deep, accurate comprehension of the physical properties involved.

While the point can be made that “true beliefs tend to be more practically useful than false beliefs” (Williamson 2020, p.15), and so we should expect our various senses and the appearance of experience to track truth because it tends to be useful, this is only true insofar as truth aligns with usefulness. In cases where something is false or a degree of accuracy is sacrificed yet is more useful or adaptive than the truth, we should

expect it to be selected for. We should not expect unuseful truths to be selected by evolution. Instead, we should expect useful falsehoods or shortcuts (or the generation of falsehoods as a byproduct of something useful) to persist. For example, a relatively simple representation of something in the environment that captures the features of relevance while sacrificing some other features or degree of accuracy could well be selected for over a more precise but resource-intensive representation.

Consider frogs who catch and eat flying flies, yet will starve when surrounded by dead flies (Mazur 1983). They have evolved a perception-action system for catching and eating flies that is based on useful information in their environment (the movement of flies in the air), yet they are tracking a kind of appearance and cannot recognise the same food source under (what is to us) a slightly different scenario. The frogs' perception-action system tracks usefulness and potential food sources, not flies as such. Their experience of the world (insofar that they have one) will include moving flies presented as a target or a food source. This is a stark example of how an organism's experience is driven and structured by their interests, not the objective nature of the world.

In support of this position, evolutionary game theory models indicate that in terms of perception-driven action, a fitness-only strategy wins out against a truth strategy (Prakash *et al*. 2019 and Hoffman 2019). That is, adaptive perception wins out over veridical perception. This result arises from perceptual ambiguity, when sensory information is first assessed for truth and then for fitness (i.e. which items may be a food source), ambiguity in the fitness dimension may be overlooked. Beyond incurring

higher cognitive costs, this strategy can also lead to less fit (useful) decisions being enacted.

In summary, the things that matter are the things that *matter* for our survival; those things that can impact on us or be exploited by us. Our biological adaptation (and to an extent our cultural development) responds to these things as we encounter them in our environment. What things are really like, beyond reliable and predictable features and behaviours is not a necessary part of this picture, just as the frog cannot recognise dead flies as a food source.

Accuracy Sacrificed

As well as potentially unnecessary, more accurate representations and seemings may have been selected against if they came with increased costs. Perhaps cognitively processing something more accurately to adhere more closely to the truth of how things are (assuming there were relevant tractable features) takes more time, which might not be selected for because "…there is often considerable evolutionary pressure deriving from considerations of speed" (Churchland 1987, p.549).

There will be many risk-benefit trade-offs in play, which evolutionary processes have respond to. This results in how we respond to our environment and bodily states being adapted to a pay-off in evolutionary terms—having surviving genetic descendants—rather than accuracy *per se*. In terms of success to pass on genes some risks will be worth taking, others will not. For example, there is the postulated 'hyperactive agency detector' (Barrett and Johnson 2003): our propensity to attribute agency as the cause of events as a survival adaptation. The idea is that our ancestors were more likely to

survive to have and raise offspring, if they tended to think the rustling in a nearby bush was due to a dangerous animal rather than caused by something without agency, like the wind. Even if we assume the rustling is far more likely to be caused by the wind, the payoff matrix favours our ancestors being on their guard and responsive to the possible threat of a dangerous agent. Fleeing from a bush rustling in the wind has relatively little cost beyond embarrassment and a few calories. Not fleeing from (or being on guard for) a dangerous agent like a tiger or hostile conspecific rival, because you mistakenly assumed the rustling was caused by the wind, has huge potential cost in terms of reproductive success. Getting eaten by a tiger considerably reduces your chance to breed successfully. Hence, given our senses are not potent enough to always reliably identify the causes of events like rustling bushes, evolutionary processes are likely to favour a bias in favour of avoiding hungry tigers and angry rivals, rather than responding at their true rate of occurrence. And so, we have bias for a sense for the presence of an agent doing agential things; usefulness over accuracy.

Being somewhat risk averse, scared of things like the dark and rustling bushes, and over-attributing agency is useful in evolutionary terms and timescales, and it is this kind of usefulness that drives how we perceive and respond to the world. It may well be why as a species we seem inclined to believe in things like gods influencing the weather, ghosts rattling pipes, conspiracy theories, and Boggarts making the milk go sour. We attribute agency too readily, because in our evolutionary past it was a worthwhile rule of thumb or strategy.

While not as vivid as the *redness* or *painfulness* that we experience, the attribution of agency is often not a merely theoretical thing or judgement. People experience a 'sense of presence' that they attribute to entities gods, demons, and spirits of the dead, and have feelings of being watched, or that someone is out to get them and the like, absent the presence of any actual agent. If one does not believe these things exist, yet people have subjective experiences of their presence, then it would in at least some cases seem to be some kind of illusionary experience.

As another example, take our 'Hyperactive Face Detector'. We over-detect faces. It is part of our tendency towards 'pareidolia'—illusory sensory perception as our perceptual system tries to detect signals amongst noise in the environment—with face pareidolia being perhaps the most recognisable example (Alais *et al*. 2021 and Liu *et al*. 2014). Many of us see faces in clouds, bushes, or even on toast. Our over-detection of faces is so natural, that we often do not notice it even while making use of it. This over detection of faces is not a case of thinking something looks a bit like a face, or an error in judgement, rather we automatically recognise it as a face, and recent research indicates that the same brain mechanisms are involved as when we see 'real' faces (Alais *et al*. 2021 and Liu *et al*. 2014).

Take my favourite illusion (Figure 3), the smiley face. Just two dots and an arc arranged just-so triggers our brains to see a face, because it is useful to spot faces, and our evolutionary environment and the risk-benefit trade-offs has made it so that we tend to over-detect faces. Absent some irregularity (such as prosopagnosia), or trauma to the brain, we do not fail to detect human faces under normal conditions. We are rather

good at it even when parts of a face are covered, appear in unusual locations, are upside down, have paint on them, and so on. But we also see faces all over the place. Figure 3 is not of a face. It does not even look like a face, not really. No real faces look like that. Yet when we look at it, we have a sense, a *feel* of *faceness*. More than that, we may get a sense of happiness. I even start to feel a little happy when looking at it. It makes me want to smile.

Figure 3 – Not a face

We cannot choose to start or stop seeing this rotated colon and closing parenthesis (Figure 3) as a face, even though it is not a face. It is not a decision we have made or that is available to us. It just seems that way to us. The likely reason why we over detect faces is that it is useful for us to detect faces, and it is better (in terms of evolutionary fitness) to over-detect faces (false alarms) than to under detect them (misses). Hence, our detection of faces works with relatively minimal stimuli, and is calibrated to be useful as opposed to accurate *per se*.

What is your brain doing now?

This is a striking question: what is your brain doing now? Though we experience things and have introspective access to our minds, we do not have access to the reality of

what our brains are doing. I suspect like me you can talk about certain kinds of processes and activities because you know at least a few things about brains and what they do, like controlling and managing your body. But this understanding is all largely sourced externally, from lectures, textbooks, journal papers and the like. Little of this knowledge comes to us from introspection. They are facts we have learned from philosophy and science, sometimes in opposition to how things seem to us. I might say that my brain is processing various stimuli and creating a sense of hunger, but I can't tell this directly. I have a slightly uncomfortable sensation that seems to originate from my tummy but going via introspection there are no signs that my brain is involved in this sense of hunger.

We do get something from introspection of course, and we have our subjective experiences, but these are rather abstracted away from what is going on. What we get are like executive reports, summations of relevant things in an easily digestible form, but only tracking the underlying data and activity rather than being directly about them. The brain is doing a whole lot of things, but what we have conscious experience of is only a small part of the overall activity, and we get something like the end results of that processing, rather than the underlying content or data. Of course, nobody really claims that we can introspect brain or cognitive processes, and the realist would claim that while we cannot access the underlying processes, we nevertheless do have access to our experiences and their properties. Afterall, they're what we are supposed to be conscious of. However, illusionists are not arguing that we do not have experiences, they're arguing that we are mistaken about what they <u>are</u>, and we are also mistaken about some of our theoretical concepts—such as the phenomenal properties—which

are not warranted. Illusionists argue that we are misled by the appearance of experiences into thinking that experiences have a certain kind of non-physical reality. But we should consider what we get in terms of the appearance of experience has come to us via evolution to be useful, not to show their reality.

Limited Perceptual Access – The Blind Brain

There is a lot going on in our brains, and we are blind to much of it. What we get is driven by at least three key factors; the evolutionary development of consciousness; its information asymmetry, and access-invariance or fixed perspective (Bakker 2012). The structure of our brain and how it processes information is the result of bottom-up process of evolutionary trial and error. What reaches consciousness is only a small part, possibly the most important part as it relates to improving our interactions with the external world. But even so, what we get is only a small part of the whole, a summarised view, weighted for understanding and action. It is not clear why, as evolved beings, we would have anything different, particularly any 'true' or 'deep' access to what's going on in our brains. When we feel fear when something jumps out at us, we do not get a sense of the various processes and low-level judgements that occurred to lead to us feeling fear. We do not get the workings; we just get the output. We just get the fear. We can only retrospectively examine what happened in theoretical terms based on feelings and perceptions that were part of our conscious experience or assess the external stimuli that we believe were present just before we felt fear. Why did we feel fear rather than surprise or joy? How close was the call between fear and joy? Would we have had a different response if we heard a friendly

greeting a second before the event, or if the lighting were better, or if we had eaten chocolate ten minutes ago, or if we hadn't? We just do not know. We do not have that kind of access to what happens in our brains. Worse, we can often be wrong about why we responded in a certain way, as is the case with confabulation (see the first paper in this thesis).

What gets to consciousness is limited by the structure and operation of the brain. There are things we know the brain does, which we are unable to get reports from, about, or influence (at least not without establishing some control loop that extends into the external world). And for that which does reach our consciousness, much of its origins are cognitively opaque to us. We are introspectively blind to how our brain works. Our thoughts and feelings come out of the darkness, we only know of them when they reach the light of consciousness, but that light gives us a limited and fixed perspective. We cannot access how our thoughts or experiences come to us, and in the main we have little influence on them. We can carefully plan what we will say next, but often the words just come. We might set ourselves a general goal for what we'll say, but often we will not know exactly what words we will use until we use them.

Much of this may be driven by the modularity of a range of activities that the brain carries out. The functional activity of at least some parts of the brain is not directly accessible to other parts of the brain. This modularity includes both encapsulation and inaccessibility (Robbins 2017). In essence much of what our brain does is carried out by processing 'black boxes', restricting and limiting access to information, with no insight from outside as to what happens within. The information flow into the process is

encapsulated, in that it is limited in what information it has access to, likely just the information inputs it receives, and information stored within the 'box' itself. The inaccessibility of modular processing is the reverse case; only the output from the 'box' is available to other processing systems, with no insight as to how the output was reached, and likely in most cases no 'meta data' about those outputs, such as how much confidence or accuracy there is in the output. It likely gives rise to range of phenomena, such as confabulation (see the first paper in this thesis).

What constitutes our conscious experiences is based on limited and heavily processed information, blind to much of the other goings-on in our brain, represented to us in a format useful to support a computational and time efficient understanding and improved interaction outcomes. It is skewed by heuristics and shortcuts for survival of genes and organisms over evolutionary timescales, driven by evolutionary trial and error. We should not expect our conscious experience to be an accurate portrayal of how things really are, but we can and should expect what we have received via these evolutionary processes is a conscious experience that is *useful* for the environmental niches it evolved in response to.

Real Patterns

Our perception of the world is shaped by evolutionary and culturally influenced responses to patterns. We detect patterns in what we observe. We're good at it (think of the smiley face in Figure 3!). What we may be detecting is a statistical regularity, a coming together of our interests.

Patterns are observer relevant. A pattern is "by definition a candidate for pattern *recognition*" (Dennett 1991, p.32). While some patterns may be readily discernible to members of the same species or culture due to their shared biology, mental processes, beliefs and values, others are not and dependent on the skills and concepts that we possess. Experts can spot patterns, diagnose problems, notice interesting trends in data, appreciate how a game is going, and so on, that others cannot, and often using only sparse data. They can pick out the most computationally useful signatures to detect a pattern. Dennett (1991, p.34) makes the point that the same environment and the same stimuli can have a range of different patterns, with varying explanatory and predictive power. We will be blind to many of the patterns that are there, because we may not have the necessary sense organs or interests to detect them.

The way things seem to us, our conscious experience, is partially built out of patterns that are relevant to us. Our folk-physics, our folk-psychology, our detection of faces when seeing two dots and a line (see Figure 3), the filling-in of the blind spot, the integration of speech and lip movement to bring us an impression of speech: all these and more are built out of patterns. Our conscious experiences are (at least partially) built out of patterns that are useful for us to perceive and interact with.

Just as we seek to detect patterns, how things seem to us, in terms of perception and experience, will have developed through evolutionary trial and error to be pattern seeking. That is, the more useful, the more tractable, the more pattern-bearing the elements of what we are conscious of are, the more adaptive they would be and so be more likely to be inherited. How they evolved to appear would have occurred in step

with how we were evolving to detect patterns. In much the same way we have cognitive heuristics and biases—useful rules of thumb and shortcuts in the way we process information and make judgements—we should think there is a good chance to have equivalent 'phenomenal biases' in the way our experience seems to us. A phenomenal bias would be a kind of systematic tendency or distortion in how experiences seem to us, which serve some useful purpose. Think of how the frog experiences potential food in the motion of flies but does not experience flies as such, or our experience of colour constancy.

The User Illusion

This picture of how things seem to us as a useful but not accurate model of information sources and things to interact with makes for a benign user illusion. It is a kind of simplified model of the things we can interact with or respond to, picking out tractable patterns, in order to bring about better outcomes than we would otherwise have without it. This is like the user illusions (or 'user interfaces') deliberately constructed in technological devices, with features such as physical controls and buttons, and virtual (or 'illusory') things to interact with in a software user-interface like files and folders. They are a benign illusion constructed for the benefit of the users to bring about better interaction outcomes, usefully deploying a simplified model or set of suggested patterns for the user to engage with rather than all the complexity and detail that is actually present in the technology. Here is Dennett (2021) talking about how the way that things seem to us is likely to have been shaped to bring about better interaction (and therefore survival) outcomes:

> "Here is a question the doubters might ask themselves: couldn't evolution have found some clever ways of installing beneficial *user illusions* in organisms that would enable them to respond under time pressure to patterns, to environmental challenges and opportunities of all sorts, dealing with a macroscopic behavioural world that was tracked by simplifications of their evolved imaginations, not a metaphysically or scientifically accurate depiction of how physical things really are? They would be the beneficiaries of these arrangements without having to understand them at all. Sellars' *manifest image*, the world we live in, is not presented to us as clouds of colourless particles, but as clumps of coloured solids, wet liquids, and invisible gusts of air and other gases, for instance. The colours that exist in the world, then, are a sort of illusion. We are *indirectly* but robustly acquainted with those properties of things in the manifest image, but the conviction that we are *directly* acquainted with 'phenomenal properties' is a confused theorists' illusion about the benign first illusion".

In the case of software user interfaces, the user illusion provides artificially constructed patterns to promote a useful understanding and prompt interaction, with the underlying workings designed to conform to the user illusion and vice versa. In the case of our interaction with the natural world the 'real patterns' are those useful detectable patterns in the external world. In the case of ourselves and each other (and possibly other species we have been close to over evolutionary timescales) there may be processes of evolutionary conformance to weak or even false patterns. That is, if assuming certain patterns or theoretical entities in others proves useful for cooperative

interaction and supports the success of gene-lines, then it may be advantageous for individuals to begin acting as though they embody those patterns or entities. This could ultimately lead to the emergence of real patterns[41]. For example, just as we see *faceness* (as much a candidate to be a phenomenal property as *redness*) in two dots and an arc[42], we also perceive it in other creatures, such as dogs. This appearance influences our interactions, prompting us to respond to them differently than we would otherwise. Through co-evolution with dogs (Schleidt & Shalter, 2003), we have developed mutually beneficial patterns of behaviour and pattern recognition. This includes our ability to read each other's body language and dogs' capability to use emotional information from human expressions (Albuquerque & Resende, 2022). It is plausible that we have even evolved to form expressions that are more easily interpretable by dogs. It is similarly plausible that the elements of our experience such as *redness* and *painfulness*, appear as they are, to guide our behaviour and interactions with others just as with *faceness*.

4.2.2 We Find our Subjective Experiences to be Useful Rather than Accurate

So far, I have made a case that as evolved organisms we should expect the way our subjective experiences seem to us to be useful rather than accurate. The way things are represented to us should show signs of being a useful model (a user illusion), rather than an accurate representation of how things are. I contend, that this is what we find.

[41] Or, putting it another way, to make that pattern more robust and reliable for use.
[42] ☺

In doing so I draw an analogy between perceptual illusions and phenomenal ones. Perceptual illusions and misrepresentations lend support to illusionism, as the phenomenal properties we seem to experience arise from our perceptions, which can sometimes be misleading or illusory. Our perceptual experiences do not always correspond with features of the external world (such as when we experience visual illusions), and even basic features like shape, length, or colour can be misrepresented. Similarly, phenomena like phantom limb pain, where the experience of painfulness is so mislocated the feeling appears to be outside the bounds of the body. These cases suggest that there is some degree of misrepresentation or illusion involved in phenomenal consciousness. In what follows I sketch out some cases which I believe support this position.

First, I want to start with a slightly unusual example of a phenomenal property, to move away from intuitions and psychological leanings that make it seem like properties such as *redness* and *painfulness* are real. Recall Figure 3, of two dots and an arc, which gives rise to the impression of a 'smiley face'. At best, the marks on the paper are a caricature or simplified model of a smiley face, providing key minimal details which are enough to suggest a face is present to our visual system. It is useful for us to spot faces, and so we are adapted to be able to do so, but clearly in the case of Figure 3 we are wrong. There is no face, just some minimal marks on a page. However, we experience the *faceness* of the marks, an experience arising from its component parts. There is something it is like to experience the *faceness* of the smiling face. It seems as though this experience is special in some way, as per other claimed phenomenal properties. For example, though some of the marks making up the smiley could be taken away, it

seems like the sense of *faceness* is simple. There does not seem to be anything that the sense of *faceness* is made up of. It isn't the case that as features are taken away there is less of a sense of *faceness*, it is there, or it is not. It seems ineffable too; it is not something I can describe and give a sense of without reference or analogy to similar experiences. Yet, there is no face, and certainly the marks on the page do not have a literal property of *faceness*, this is something that is added to the picture via how it is represented to as an experience. In the case of *faceness*, there is less of a sense that the *feel* may be distinct from any representational or functional role, as compared to *redness* or *painfulness*. My view is that this should make us reappraise how we think about 'classical' phenomenal properties, rather than deny *faceness* a similar standing to the likes of *redness* and *painfulness*.

Consider also our experience of finding out about ourselves, of reporting our propositional attitudes like our beliefs, desires, and intentions. We seem to experience our first-person knowledge as a kind of direct and transparent access, but there is substantial evidence including the occurrence of confabulations which suggests that it is not, and our access may be interpretative in nature instead (for example, see Carruthers 2013, Gopnik 1993, and the first paper in this thesis). The experience of our first-person knowledge as direct and transparent may be a kind of illusion (Gopnik 1993), arising from an inbuilt transparency assumption (Carruthers 2013, pp.11-18). If so, the illusion may be useful in fostering a sense of confidence and unity in our psychology.

Next, consider again the case of phantom limb pain. The location of a *pain* experience is an inseparable part of the experience. We do not get free-floating *pain*. When we experience *pain*, something *hurts*. The spatial nature of it is part of the *pain* experience, which cannot be separated out. Yet, we have experiences that seem as if they involve *painfulness* in bodily locations that do not exist. If one of our legs is missing and we feel a *painful* throbbing sensation in it as though someone had just kicked us in the missing limb, then there is a phenomenal misrepresentation in play.

Not only is there a phenomenal misrepresentation in where the *pain* experience is, that the *pain* has a bodily location at all is a (useful) misrepresentation. The experience of pain is that it occurs in the body. However, the strong evidence for neural correlates of *pain* (Ridder *et al.* 2021), the existence of phantom limb pain, and being able to stimulate the brain to bring about experiences of *painfulness* in the body (Labrakakis 2023) tell us that *painfulness* occurs in the brain rather than being located in the body. So not only can we have misplaced *pains*, feeling *pain* in the wrong areas or in missing parts of our body, that we feel bodily-*pain* at all is a useful misrepresentation.

Further, if we are feeling *pain* in a missing limb, not only is there a misrepresentation in the location of the *pain*, there is also a misrepresentation in the character of the *pain*. We cannot feel *pain* in a limb that does not exist, so the presence of the *pain*, the kind of *pain* it is (throbbing, burning, stabbing, etc.) cannot correspond to the reality of the pain in the missing limb. More generally, what it is like to feel *pain* tends not to track the underlying pain event and is moderated by various affective factors. Rather than a linear response of increasing feelings of *pain* as a pain-stimulus increases, what we get

for our feeling of *pain* is a sigmoid function (Borstad 2015). Initially, as the stimulus increases, we have no or limited feelings of *pain*, then as the stimulus increases further our feeling of *pain* increases sharply, before eventually substantially reducing in the rate of increase. The *feel* of pain does not track, does not accurately represent, the pain-causing stimulus, rather it represents it in a way to provoke a behaviourally useful response.

These cases show that phenomenal properties of our subjective experiences can be misrepresentations and are therefore in those cases illusory. To be sure, I have not made a case that all aspects of experiences are illusory, but I have shown that at least in part they can be illusory. To add to this, next I make a brief case drawing on perceptual issues that we are inclined or predisposed to overlook or not notice instances where our representations are not accurate.

We find it hard to notice the lack of accuracy because the shortcomings are omitted

As previously mentioned, when discussing accepting appearance-reality gaps, we have a visual blind spot. It is always present, but we cannot see it. Our brain compensates for this blind spot caused by filling in the gap in our perception. As a result, we are blind to our blindness and experience a whole visual field, with the world appearing to us as it is not. We cannot discover this filled in gap or notice it through introspection alone, we need a change of perspective to be able to find it. It is not like we see a hole in our vision; we have no sense or perception of the hole. It does not perceptually exist for us, even as an absence.

This absence, and the absence of any indication that there is an absence, is not an inevitable outcome following on from the architecture of the eye. It could have been otherwise. There are other options available in the 'design space' of how the blind spot could have appeared to us. For one, it could have been a fuzzy patch in our vision, for another we could have had a sense of absence about that spot in our visual field, much like we might have a sense of *faceness* or *happiness* from seeing a smiling face. But that is not the way it is. The appearance of the phenomenal visual field conceals the reality of the absence from us, and presents us with a different phenomenal view of the world

In other ways our perception of the world is not as complete as we think it is, based on how things seem to us from an uninformed perspective. Our eyes can only see detail in a small portion of our field of view at a time, with the periphery of our vision being more sensitive to movement than colour or detail. Yet this conflicts with what seems to us like a full detailed view from our perspective. In what Noë (2002) calls the 'Grand Illusion', rather than having a rich view of the world, or even a detailed mental model of a visual scene to reference, we instead often access visual information from the world in a just-in-time on-demand fashion to get the information as it is needed. Our perceptual experiences mislead us as to what we see and how we see the world. Instead of experiencing the reality of what is perceptually available to us, we have an appearance of a much richer and more detailed visual field.

As mentioned earlier in this paper, there are all kinds of quirks and kludges and guessing going in our sensory experiences. For example, our perception of colour remains relatively constant from different viewing angles and lighting conditions, even

though what reaches our eyes will be quite different (Foster 2011). The McGurk effect (McGurk and MacDonald 1976), where the bringing together of sight and hearing of speech can lead to a perception of things being said that are not being said, demonstrates the integration of different sensory modalities and the guesswork that goes into linking heard and viewed speech.

To take another case, our vision comes to us as though in real-time without any breaks in it, somewhat like a movie. Or so it seems. The reality is our visual feed has interruptions in it. We do not perceive visually what happens when our eyes jump around in a saccade (Irwin *et al.* 1998) when we are looking around. We are blind to the interruptions, blind to the fact that for a brief span of time we are not able to see.

These phenomena extend to our internal senses too. It seems that excitement and anxiety can be mistaken or made to be the other depending on our attitudes and the context (Brooks 2014). What are the same signals, the same root sensations, become very different things. The misattribution error shows the role our experiences, thoughts and concepts, shape how our feelings are interpreted. In the shaky bridge experiment (Dutton and Aaron 1974) males were more likely to try and get contact details from a female researcher after they crossed a 'shaky bridge' than in a more sedate location. This research points to a misattribution of the cause of their emotionally aroused state from the fear-inducing bridge to the attractiveness of the person in front of them. This shows that sometimes we do not even know what emotion we are feeling, and what we experience is shaped by other factors such as judgements about our situation (see also, my paper on Confabulation in this volume).

In short, the way things seem to us is often misleading if what we care about is truth or accuracy. We should expect this to be the case, as we are evolutionary adapted to survive and pass on or help aid the survival of our genes. We can see a wide range of evidence that this is the case when it comes to perceptual shortcomings, heuristics, and illusions. I have also made a case that we do have at least some evidence for the same when it comes to subjective experience, drawing on the examples of *faceness* and phantom limb pain.

4.2.3 The Good News: it is Useful

The good news is that while our conscious experience is not fully accurate, it is useful, and it is useful because it is accurate in a certain kind of way. We may not get 'the true picture' of reality, but what we do get is usefully calibrated to evolutionarily useful risk-benefit trade-offs, picking out tractable and useful real patterns. We do not perceive the world with complete accuracy, as is shown by the various illusions we are subject to. Yet, our senses and how thing appear to us are aligned towards usefulness. Take the hyperactive agency detector idea; we're wired to detect agents (like predators) quickly, even if it sometimes leads to false alarms. The brain processes involved track features that are detectable and relevant to us, rather than the truth of the matter. As an evolved capacity, we should consider that this applies to phenomenal consciousness as well, as even "distorted representations may still carry useful information" (Frankish 2023, p.7), and sometimes distorted ones may bring out or emphasise the salient features and minimise or eliminate less relevant features, such as with a cartoon or caricature, to be more effective than a 'truer' representation.

The illusion of phenomenal consciousness may not accurately represent what things are really like, but it represents and reliably tracks features and behaviours that are relevant to us, the things we use to distinguish between objects, interact with them, and respond to them. So, we see vibrant and dank colours when appraising objects. We perceive affordances, with things that are touchable, graspable, turnable, zippable, and the like, jumping out at us.

Consider a map as an analogy to phenomenal consciousness. A map shows a geographic area but does not replicate the area exactly. More than that, many maps are not good representations of the reality of the geographic area. They have features on them like coloured lines to mark out different types of roads, and symbols to mark out features of interest. In terms of accuracy, maps are poor representations, they're nothing like the reality. Roads are much smaller than they appear on maps and they are not brightly coloured. Features of interest do not look like they do on the map, and contour lines cannot be seen on the ground. In terms of accuracy, maps misrepresent the reality. However, in terms of usefulness maps can be excellent representations. The road lines on the map reliably track real features in the world and tell us relevant and useful information about the categories of roads. Markings on maps make features of potential interest salient and easy to locate. Contour lines are useful in helping to plan out routes and avoid dangerous slopes. Phenomenal consciousness is like this, in being useful rather than accurate. If we could only ever navigate by a map and not access the reality of the world that the map depicts, then we would be greatly misled as to what the world was really like. At the same time, if the map was useful enough and it is all

we knew, we might find it difficult to believe that we were dealing with an appearance and not the reality.

The appearance, the form, of our conscious experience—our user-illusion—is a useful and reliable guide to things that matter to us in the environment. And when you have a good model like this, you may 'mistake the map for the territory', mistaking the model for the reality, which is all too easy to do in the case of conscious experience because we cannot get a direct and unmediated view of the reality. The map of the territory is all that we have. Psychologically the illusion and reality collapse into one.

A reason why we find it counter-intuitive to accept we are interacting via a user-illusion is that it is rather good. It is useful. It does make salient useful features and affordances in our environment. It does reliably track changes that matter to us. This is why it can be difficult to accept our conscious experience is not a direct reflection of reality, but a construction shaped for our survival. That, and the discrepancies and illusions that we do encounter (such as the blind spot or having a sense of *faceness* from a colon and an arc (see Figure 3)) do not stand out. We are blind to the blind spot as the existence of the absence is itself absent from our visual perception. We are not directly aware of flaws in the illusions, or they seem natural and normal to us, such as accepting Figure 3 has *faceness* as though it wasn't a marvellous exploitation of our visual processing system.

4.3 Summary

I have set out to show how we should not expect our phenomenal consciousness to be an accurate guide to what our phenomenal consciousness is really like, or what its

content is like. Usefulness over accuracy is what we should expect of evolved cognitive systems.

While I have not been able to directly tackle phenomenal *feels* like *redness* or *painfulness* head-on, I have sketched out how the way things seem to us is generally a sufficient and useful user-illusion laden with patterns. An analogous case is of '*faceness'*, a sub-personal evocation of a face from sparse lines on a page, an illusion that we cannot help but be subject to. We know there is no face present, but even so get the sense of *faceness*. This is a basic example to be sure, but one can (perhaps) begin to see how other *feels* like *redness* and *painfulness*, can be similarly illusory. Even if this approach is not sufficiently convincing, it does undermine the grounds to accept phenomenal consciousness as something that does not fit into the physicalist picture of the world.

5 The Defence

Illusionism is a tricky thesis to argue against. It argues that the way phenomenal consciousness—the *feels* like *redness* and *painfulness*—seems to us is not how it really is, that there is an appearance-reality gap, and what how it seems to us is an illusion. Illusions have a reality, but it is a reality of appearance, not of how things are, much like the case with rainbows. To fully refute illusionism, an argument needs to make the case that our phenomenal consciousness is not or could not be illusionary.

What follows is a consideration of three critiques against illusionism, which can also be levelled against the view I have given in this paper. The first critique denies there is a gap between appearance and reality, holding to the position that our experience of phenomenal consciousness is the reality. This preserves the appearance and denies that it is illusory but does not in and of itself preserve the phenomenal features.

The second argument against illusionism is to embrace the primacy or authority of our experience of phenomenal consciousness. The move here is to say that whatever else may be the case we can be sure that we are having phenomenal experiences, and so anything that casts doubt on that must give way to the primacy of our immediate experience. This makes intuitive sense, and is related to Chalmers' (1996, p.188) *datum* argument discussed earlier (in Section 3.2.1); it is taking what we experience as primary. Adopting this approach, one could argue that our default assumption should be to take appearances at face value, but this position comes under strain given we know how thing appear to us can be misleading, and as per Frankish's argument

against anomalousness (see Section 3.2.2), we should be suspicious of things that do not fit into the physicalist worldview.

Finally, the last argument against illusionism considered here is a critique that in accepting the illusionist position, we would be denying or reducing people's moral worth. The idea is that if *painfulness* is merely an illusion, then we are less obligated to others. It remains however that people do not like whatever it is that is happening when we talk about experiencing pain, and so there remains a foundation for caring about the welfare of others, even if we were to reject the reality of their subjective experience of pain.

5.1 No Appearance-Reality Gap

A critic of my position and illusionism more generally can argue that there can be no appearance-reality distinction in the case of consciousness (Shabasson 2021). The idea is that we cannot have a dissociation between appearance and reality for conscious experience in the way that it is possible with ordinary perception. If I seem to be perceiving an apple and have an experience of *redness*, then I might be mistaken about seeing an apple—I could be hallucinating or seeing some kind of illusion—but I cannot be mistaken about having the experience of *redness*. If I have an experience of *redness*, then it does not matter whether there is an appropriate association with a referent or not, I have experienced *redness*. An illusion of an experience would necessarily be an experience (Strawson 2018).

I agree that there is something special in the case of consciousness, that with respect to experience it is the appearance that matters. We might be mistaken about what we

have experienced and what the experience is of, but if we think we're experiencing something then it seems impossible for us to not be having an experience. However, this does not undermine there being a gap between that which is representing and what is being represented. This is similarly the case when experiencing a simulation, watching a film, or playing a computer game. The appearance is the relevant reality for the experience, but that says nothing about the reality underlying the appearance.

I cover two different but related arguments against there being an appearance-reality gap; the reference argument, and the idea that an illusion of an experience is still experienced.

5.1.1 The Reference Argument

Prinz (2017, p193) argues against illusionism, via the role of quasi-phenomenal states. These states are not phenomenal but are (mis)represented as such and are one suggested mechanism for how the illusion of consciousness comes about (Frankish 2017, p.15-16). Prinz (2017, p.193) argues that when we talk about phenomenal states we are referring to those inner events that seem to have phenomenal features (e.g. *redness* or *painfulness*), and therefore quasi-phenomenal states (if they exist) are phenomenal states.

The general thrust of this argument is that we really do seem to have these experiences which include phenomenal aspects such as *painfulness*, and so whatever they are or however they work they are phenomenal. Whether there is an illusion

involved or not, there is the *painfulness* of pain, and so we must have phenomenal states.

One way to make sense of this view is by analogy: imagine we see something in the sky and Bloggs says it is a light, and Smith says it is a planet which is reflecting light. On further investigation we find that the latter is true; it is a planet reflecting light from the sun. We have increased our knowledge. We know more about the light in the sky. But on this view, we have not proven Bloggs wrong. The phenomenon that Bloggs was referring to still exists. We still see a light in the sky. The same is the case for phenomenal states even if Frankish is right about quasi-phenomenal states. What Prinz is saying, is that when we talk about phenomenal states or quasi-phenomenal states we are talking about the same thing, because we are pointing to the same set of phenomena. For Prinz, even if Frankish is successful in changing our understanding of conscious experience, there is still something performing the role of what we refer to when we talk about phenomenal consciousness (there is still a light in the sky), and so illusionism collapses into realism. Or realism inflates into illusionism.

However, what the illusionist can say is that while Bloggs is right that there is a light in the sky, it is not the source of the light, it is 'just' a reflection. It *seems* like a light source, and from various perspectives it might be indistinguishable from a true light source. But, as we learn more about it, we might come to understand that whatever it is, it is not the sort of thing that emits light like we first supposed. We would have something anomalous to our understanding of how celestial bodies and light emission works. Upon further investigation we would discover that it is a reflector of light rather

than a source of light. The same can be said of conscious experience; it seems a certain way—to have *redness* and *painfulness* and so on—but they're not real things, it only seems that way due to our limited perspective on natural processes, giving rise to an illusion. Where the analogy departs is that in the case of the reflecting planet, there really is a true light source. Sure, somewhat different, and in a different location, but the same sort of thing as what we were expecting the planet to be. Illusionists would say that in the case of phenomenal consciousness our misrepresentations of the relevant properties and states are so 'distorting' that they appear to be something completely different to the reality.

Frankish (2017b, p283) responds to Prinz's line of argument by saying that his strategy would work if we were pointing to the same target, but we are also describing the target. If our set of descriptions do not fit or are in conflict, then we may not be able to say they are of the same thing. Frankish says that if our descriptions of phenomenal states are radically different to what they are supposed to be describing, then we shouldn't continue to use those descriptions. The realist could reply that (in the main) our descriptions are of the experience, which the illusionist isn't denying, and so the descriptions aren't radically different. This lines up neatly with the example of the light in the sky. It depends on what Bloggs is saying about it. If he is merely pointing to and describing a light in his visual field, then suggestions that it is an illusion are misleading and not relevant. If, however, he talks about it as though it were the source of the light, or of the object having some kind of light emitting or generating capacity (perhaps theorising about energy sources on the planet for example) then we would be entitled to say that he is mistaken, and it only seems as though there is a source of light in the

sky. The same holds true for experiences like *redness*, *painfulness*, and *faceness*; so long as we stick to how they appear, we are not making any theoretical errors, they are like the lights we see in the sky. If we start attributing theoretical concepts to them, like phenomenal properties—or making claims as to whether they are a source of light or a reflection—we are making a mistake.

5.1.2 An Illusion of an Experience is Necessarily an Experience

Some argue (e.g. Strawson 2018) that there is no appearance-reality gap when it comes to phenomenal consciousness, that an illusion of an experience is still necessarily an actual experience. This parallels Prinz's reference argument above, in effect saying whatever is going on, there is something that gives us an experience, whether the experience is of an illusion or not, and that given we have experiences which are described as phenomenal, that which gives rise to them must therefore have phenomenal properties (Pereboom 2017).

Whether this succeeds hinges on what we mean by having an experience, what it is *of*. When I see a red stripe in the world or feel a pain, where are they to be found? The illusionist position is that when we see red or feel a pain, there is no *redness* and there is no *painfulness* to be found. *Redness* and *painfulness* are not physical things. They're not out there in the world to see or feel. Nor are they in the brain. The experience of red or pain is a representation or judgement about internal states, but there is no further existence to *redness* or *painfulness*.

The distinction seems to be where the phenomenal feels (quasi or otherwise) are located in the causal chain between sense organ (or some other sense capability) and consciousness. Illusionists say there is no *redness* to be 'seen', that the impression of *redness* is a kind of introspective judgement about the information our brains have gathered and processed. It's the act introspection that brings about the sense of *redness*, as a kind of judgement. Our response to mental states is what gives us a sense of *redness*. But representations do not need to have the properties that they're representing (Frankish 2019), just as a map does not need to have actual hills or roads, and we wouldn't expect something in our brains or minds to become *weighty* when we heft a dense object. In contrast, the phenomenal realist position seems to be that there is *redness* to be introspected. So, we have *redness* that can be introspected, or an act of introspection that results in a sense of *redness*, but just as there isn't anything *weighty* in an experience there is not any *redness*, they're a kind of judgement.

5.2 The Primacy of Conscious Experience

First-Person Authority…

Balog considers illusionism to be an extraordinarily implausible position (Balog 2017 p47). She accepts that introspection might provide us with misrepresentations, such as illusions, and rightly suggests that we have no way of knowing whether that is the case in any given instance (Balog 2017 p47). However, Balog thinks there is no reason to reject the existence of phenomenal consciousness and considers the debate to be one between the epistemic authority of first-person experience versus scientific (and presumably philosophical) theorising (Balog 2017 p47).

Balog considers illusionism to be in conflict with the fundamental way in which we perceive the world and ourselves (Balog 2017, p47). Balog (2017) says that our awareness of phenomenal consciousness is a "bedrock feature of what it is to be a human being" (p47); there are aspects of our lives which are ineffable and closed off to scientific investigation (p42); a purely objective complete account of ourselves would leave out things like value and meaning (p42); and Balog is prepared to reject physicalism before she rejects first-person authority about phenomenal matters (p43). The general thrust of Balog's position seems to be that theorising based on an objective investigation of the world is too limited to ever be able to overturn our first-person experiences.

…is not immune to illusions

However, illusionism does not overturn our first-person experiences. It softly endorses them; yes, we do have this experience and all its attendant secondary impressions and responses, but no, we are wrong about what it *is* that we have experienced. Because Balog's argument against illusionism is based on the compellingness of phenomenally conscious states, she is arguing for the *strength* of the experience but not against it being an illusion. Illusions need not be weak, uncompelling, or seem inauthentic!

If one thinks one is seeing a real red apple, but is told that it is an illusion, then responding that no, you really can see a real apple and you can tell that it is real because you can see it does nothing to prove your position. If the apple *is* an illusion, your protests prove that the illusion is convincing. To resolve whether it is an illusion or not, the object and process that leads to the perception of an apple needs to be

investigated. Perhaps a realist sharing Balog's views could say that as we cannot investigate and check our introspections, we cannot prove the case either way so we should stick with first person authority. I think that, in general, we can carry out such an investigation, and that the known flaws with introspection give us ample reason to think it may mislead us in other ways. If introspection systematically misrepresents content to appear a certain way, it will be difficult or impossible to see otherwise via introspection. That the brain and brain processes seem *prima facie* unrelated to our thoughts and feelings, and yet we know are vital to them, is evidence for how challenging we consider thinking of such things in terms of their causal mechanisms.

A move that someone backing our first-person authority about subjective experiences can make is to place that experience front and centre, and say whatever it is, whether empirical evidence may count against it, it is real because it has a reality to us. One route to take this might be a form of idealism, but another is to take the appearance at face value but insist that it is only an *appearance*, that is to endorse illusionism.

5.3 Throwing the Baby out with the Bathwater

Strawson (2018) thinks that illusionism, which he considers to be a denial of consciousness, is the silliest claim ever made. So silly a claim, that he calls it The Great Silliness. Strawson (2018) thinks the denialists (i.e. illusionists) deny that anyone has ever had an experience of seeing, hearing, or smelling something, or of feeling pain, hunger, hot or cold, or any emotion, or even of feeling sleepy. Essentially, Strawson thinks that illusionists are denying that we have any experience at all. Strawson thinks this denial arises from the mistaken view that what subjective experience seems to be

like (i.e. the featuring phenomenal properties) is incompatible with physicalism. He thinks that the illusionist project is arguing to do away with subjective experience; in essence, throwing the baby out with the bathwater.

Strawson (2018) is an "outright realist" about subjective experience. To Strawson phenomenal consciousness is obviously natural, because it cannot be anything else. But he strongly endorses how phenomenal consciousness seems as indicative of its nature, rejecting illusionism. Strawson accepts that 'the deniers' do not think they are denying subjective experiences but thinks that in denying the phenomenal properties they are (whether they intend to or not) denying experiences. To Strawson, the illusionist project is not merely reductive, it is eliminative, as he thinks that what illusionists say about what subjective experience really is cannot be compatible with how it seems to us.

Strawson thinks this disagreement about subjective experience has substantial implications. He thinks that without the *painfulness* of pain, we cannot say that someone has suffered, and this raises substantial moral issues with regards to perceived suffering in the world (Strawson 2018). Further, if nobody has ever felt *pain*, then nobody has ever caused *pain*. If nobody has ever caused someone else *pain*, then perhaps it undermines our practices of condemning people and holding them accountable. The worry here is that if we only seem to experience *pain*, but that *pain* is just an illusion, then we're a kind of sophisticated robot that only seems to feel pain when we get kicked in the leg. This would undermine our status as moral subjects.

There are two things to say in response to this worry, the first is that it is not clear that Strawson is correct to suggest that absent the *painfulness* of pain that there is a substantial moral issue. Neil Levy (2024) argues that while having phenomenal consciousness may be sufficient to have moral value, it might not be necessary. Levy considers his subjective experiences to be relatively impoverished in comparison to how others talk about and ascribe value to theirs. Levy (2024, p. 20) insists he is not a zombie, but even with his impoverished experiences, it would seem a leap to say that there were different moral issues as to how we should regard his suffering. Furthermore, it is clear that we do not want to experience pain, and various moral ailments inflicted on us will harm or hamper our interests. This may be sufficient to ground moral considerations in functional properties.

The second response is that the illusionist project does not deny pain or suffering anyway, it says something about what they are, what the experience is. Contrary to Strawson's concerns that illusionism is an eliminativist project, illusionism preserves the phenomena of consciousness, albeit someone aligned with Strawson may see this as a kind of hollowing out of the phenomena. Illusionism claims that things like *redness* and *painfulness* are attributed in judgement (Frankish 2023), a set of responses to representations. Illusionists are not arguing that there is no such thing as subjective experience, they're arguing that there is a substantial difference between what it seems to be and what it actually is. It is, at best, premature to claim that illusionism undermines the value of experience, or of human beings.

To be clear, neither my view or the illusionism project claims or entails that there are no red things, or that no one feels pain. Things have red properties, and we experience the red. People feel pain. What is being denied, is theoretical commitments about what our experiences are and how they fit into the world.

Strawson (2018) acknowledges this move but thinks it is unachievable without getting rid of consciousness from the picture. He argues that to reduce consciousness to behaviour and dispositions and the like is to eliminate it, on the basis that whatever consciousness is, it is not behaviour and dispositions. I think the mistake being made here is that the illusionist move does not do away with the appearance of consciousness, just as learning that tables are mostly void at the atomic level does not do away with solid surfaces, or that lightning is electrical discharge does not do away with the appearance of lightning. Illusionism is an explanatory project; it does not seek to do away with the experience of consciousness any more than learning about biological mechanisms does away with the phenomenon of life, or learning about atoms does away with tables.

Perhaps the lingering worry is that if things like pain were 'merely' informational states and related reactive attitudes, is there really any *painfulness*? A robot built today could have informational states registering that insult or injury has been inflicted on it when you kick it in the shin. The robot could also have a whole set of secondary responses to these informational states, including behaviours and further representations like broadcasting its pain to others through making a yelping noise, and an aversion to the shin-kicker. The claimed intuition is that the robot has not experience *painfulness*. It

seems like something further is missing. Surely there is something else to pain? Well yes, that is likely the case if a contemporary robot cannot experience *painfulness* but we can. But that does not mean it must be some further thing over and above what our brain does. If we—or a similar organism—yelp and scream in response to an event, flinch away from it, are highly motivated to avoid it happening again, and fully believe that we have had an experience of *painfulness*, it is not clear what it would mean to say the *painfulness* was absent.

The alternative, if pain is not something like an informational state and our responses to it, is that there is some further thing, a 'pain in itself' or the 'essence of pain'. However, problems arise from this. What or where is this real pain, and what it could add which is not already included in the information states and responses to them, seems like looking for the essence of life. Ultimately, whatever its nature, the experience of *painfulness* is unwelcome. Whether an illusion or not, sentient beings do not like it, seek to avoid it, and can experience debilitating psychological and physiological effects from it. Strawson would say it is the qualia that make it unwelcome, but illusionists say there is no such thing as qualia but even so the appearance of *painfulness* is still unwelcome. A mistake here is to think that in labelling something as an illusion is to say there is nothing present or that the illusion itself does not matter. Neither is true. Under illusionism we still have an appearance of *redness*, we still have an appearance of *painfulness*, just as we still see rainbows, enjoy their beauty and produce art about them, even though they do not have a real physical presence.

6 Conclusion

The apparent anomalousness of our conscious experience in a physicalist worldview can potentially be resolved in several ways. Illusionism attempts to do so by denying that the appearance of phenomenal consciousness is real, arguing that the appearance is an illusion. In support of the illusionist approach, I have given some brief challenges to the so-called Hard Problem of Consciousness, including critiquing a set of supposed second-order phenomenal properties (simple; ineffable; intrinsic; private; and immediately apprehended), giving reason to doubt that they conflict with physicalism.

My main argument—intended to help 'set the scene' for illusionism—is that we should expect our perception of the world, and more broadly our subjective experiences, to be shaped for usefulness rather than accuracy. As evolved beings, our cognitive and perceptual systems have adapted to prioritize usefulness over truth, making our perceptions reliable only to the extent that they track utility. Given the inherent fallibility and susceptibility to illusions and biases in our cognition and perception, we should not expect the way our subjective experiences seem to us to provide an accurate guide to reality. Instead, our phenomenal consciousness, which I take to be a product of evolution, likely developed not to reveal the true nature of things, but to serve as a useful guide—a sort of user illusion—helping us to interact effectively with our environment and with others. This view is reinforced by the many ways in which our experiences are demonstrably incorrect or illusory, such as our blindness to the visual blind spot and the various visual illusions to which we are subjected.

Based on what I have argued, I take the Illusionist project to be the most viable approach to responding to the Hard Problem of Consciousness, and the most scientifically conservative viewpoint to take. However, there remain issues with illusionism and attempts to explain subjective experience, not the least of which is why getting kicked in the shin feels quite so *painful*.

While I do not expect the arguments I've given here to be knock-down arguments against the Hard Problem or in favour of illusionism, my main argument more broadly construed shores up what I consider to be an epistemic virtue when it comes to considering naturally evolved systems. We should not expect perfection or a lack of compromise or a lack of quirks or shortcuts in such systems, and as such we should be cautious about giving strong endorsement to how things seem to us as a reflection of how things really are.

7 References

Alais, D., Xu, Y., Wardle, S. and Taubert, J. (2021) 'A Shared Mechanism for Facial Expression in Human Faces and Face Pareidolia', *Proceedings of the Royal Society B*, 288, no. 1954, pp.1-8.

Albuquerque, N. and Resende, B. (2022) 'Dogs Functionally Respond to and Use Emotional Information from Human Expressions', *Evolutionary Human Sciences*, 5, e2, pp.1-10.

Anscombe, G.E.M. (1965) *An Introduction to Wittgenstein's Tractatus*. Harper and Row: New York.

Bakker, R. (2012) 'The Last Magic Show: A Blind Brain Theory of the Appearance of Consciousness'. Available at: https://www.academia.edu/1502945/The_Last_Magic_Show_A_Blind_Brain_Theory_of_the_Appearance_of_Consciousness accessed: 15/01/2023.

Barrett, J. and Johnson, A. (2003) 'The Role of Control in Attributing Intentional Agency to Inanimate Objects', *Journal of Cognition and Culture*, 3.3, pp.208-217.

Balog, K. (2017) 'Illusionism's Discontent', in Frankish, K. (ed.) *Illusionism as a Theory of Consciousness*, pp 40-51. Imprint Academic: Exeter.

Blackmore, S. (2022) 'The Grand Illusion', in Symes, J. (ed.) *Philosophers on Consciousness: Talking about the Mind*. Bloomsbury Academic.

Borstad, J. (2015) 'The Role of Sensitisation in Musculoskeletal Shoulder Pain', *Brazilian Journal of Physical Therapy*, 19(4), pp.251-256.

Block, N. (1995) 'On a Confusion about a Function of Consciousness', *Behavioural and Brain Sciences*, 18, pp.227-287.

Brooks, A. (2014) Get Excited: Reappraising Pre-Performance Anxiety as Excitement', *Journal of Experimental Psychology: General*, Vol.143, No.3, pp.1144-1158.

Carruthers, P. (2013) *The Opacity of Mind*. Oxford University Press.

Chalmers, D. (1995) 'The Puzzle of Conscious Experience', *Scientific American*, 273, pp.80-86.

Chalmers, D. (1996) *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Chalmers, D. (2018) 'The Meta-Problem of Consciousness', *Journal of Consciousness Studies*, Vol.25, No.9-10, pp.6-61.

Choudhury, R. (2016) 'Chromatic Adaptation and Colour Constancy', in Kumar, A. and Choudhury, R. (eds.) *Principles of Colour and Appearance Measurement – Volume 2: Visual Measurement of Colour, Colour Comparison and Management*, pp.214-264. Woodhead Publishing. Available at: https://doi.org/10.1533/9781782423881.214. Accessed 28/12/2022.

Churchland, P. (1987) 'Epistemology in the Age of Neuroscience', *The Journal of Philosophy*, Vol. 84, No. 84, pp.544-553.

Churchland, P. (1996) 'The Hornswoggle Problem', *Journal of Consciousness Studies*, Vol.3, No.5-6, pp.402-408.

Clark, A. (2014) *Mindware*, 2nd edition. Oxford University Press.

Dennett, D. (1979) 'On the Absence of Phenomenology', in Gustafson, D.F., Tapscott, B.L. (eds.) *Body, Mind, and Method*. Synthese Library, vol 138.

Dennett, D. (1988) 'Quining Qualia', in Marcel, A. and Bisiach, E. (eds.) *Consciousness in Modern Science*, Oxford University Press. Available at: http://cogprints.org/254/1/quinqual.htm accessed 24/02/18

Dennett, D. (1991) 'Real Patterns', *The Journal of Philosophy*, Vol.88, No.1, pp.27-51.

Dennett, D. (1993) *Consciousness Explained*. Penguin.

Dennett, D. (2013) 'The Curse of the Cauliflower', *Intuition Pumps: and other tools for thinking*, pp.296-298. Allen Lane.

Dennett, D. (2020) 'A History of Qualia', *Topoi*, 39:5, pp.5-12.

Dennett, D. (2021) 'The User-Illusion of Consciousness', *Journal of Consciousness Studies*, 28, No. 11-12, pp.167-177.

Díaz, R. (2021) 'Do People Think Consciousness Poses a Hard Problem?: Empirical Evidence on the Meta-Problem of Consciousness', *Journal of Consciousness Studies*, Vol.28, No. 3-4, pp.55-75.

Dutton, D. and Aaron, A. (1974) 'Some evidence for heightened sexual attraction under conditions of high anxiety', *Journal of Personality and Social Psychology*, 30 (4), pp.510-517.

Eddington, A. (1927) 'Gifford Lectures: Introduction'. Available at: https://mathshistory.st-andrews.ac.uk/Extras/Eddington_Gifford/ accessed 14/09/2024.Fish, W. (2010) Philosophy of Perception: A Contemporary Introduction. Routledge: New York and Abingdon.

Foster, D. (2011) 'Colour Constancy', *Vision Research*, Vol.51, No.7, pp.674-700.

Frankish, K. (2017) 'Illusionism as a Theory of Consciousness', in Frankish, K. (ed.) *Illusionism as a Theory of Consciousness*, pp.11-39. Imprint Academic: Exeter.

Frankish, K. (2017b) 'Not Disillusioned – reply to commentators', in Frankish, K. (ed.) *Illusionism as a Theory of Consciousness*, pp.256-289. Imprint Academic: Exeter.

Frankish, K. (2019) 'The Consciousness Illusion', *Aeon Magazine.* Available at: https://aeon.co/essays/what-if-your-consciousness-is-an-illusion-created-by-your-brain accessed 25/01/2025.

Frankish, K. (2022) 'Like a Rainbow', *Keith Frankish* (personal website). Available at: https://www.keithfrankish.com/2022/05/like-a-rainbow/ accessed 17/09/2022.

Frankish, K. (2023) 'What is Illusionism?', *Klēsis Revue Philosophique*, 55, pp.1-17.

Goff, P. (2017) *Consciousness and Fundamental Reality*. Oxford University Press, New York.

Goff, P., Seager, W., and Allen-Hermanson, S. (2022) 'Panpsychism', in Zalta, E. (ed.) *The Stanford Encyclopaedia of Philosophy* (Summer 2022 Edition), available at: https://plato.stanford.edu/entries/panpsychism/ accessed 15/09/2022.

Gopnik, A. (1993) How We Know Our Minds: The illusion of first-person knowledge of intentionality, *Behavioural and Brain Sciences*, 16, pp.1-14.

Hall, G. (2021) 'Phenomenal Properties are Luminous Properties', *Synthese*, Vol.199, No.3-4, pp.11001-1022.

Hoffman, D. (2019) *The Case Against Reality*. Penguin.

Humphrey, N. (2012) *Soul Dust*. Quercus.

Irwin, D., Brown, J. and Sun, J. (1998) 'Visual masking and visual integration across saccadic eye movements', Journal of Experimental Psychology, 117(3), pp.276-287.

Jain, A., Bansal, R., Kumar, A., and Singh, K. (2015) 'A Comparative Study of Visual and Auditory Reaction Times on the Basis of Gender and Physical Activity Levels of Medical First Year Students', *International Journal of Applied and Basic Medical Research*, Vol 5(20), pp.124-127.

Jaworski, W. (2011) *Philosophy of Mind*. Wiley-Blackwell.

Kitaoka, A. (2012) 'Fraser-Wilcox Illusion', *Akiyoshi's Illusion Pages*. Available at: https://www.psy.ritsumei.ac.jp/akitaoka/Fraser-Wilcox-illusione.html accessed 18/06/2025.

Labrakakis, C. (2023) 'The Role of the Insular Cortex in Pain', *International Journal of Molecular Sciences*, Vol.24, No.6, pp.1-16.

Levine, J. (1983) 'Materialism and Qualia: The Explanatory Gap', *Pacific Philosophical Quarterly*, 64, pp.354-361.

Levy, N. (2024) 'Consciousness Ain't All That', *Neuroethics*, 17:21, pp.1-14.

Liu, J., Li, J., Feng, L., Li, L., Tian, J. and Lee, K. (2014) 'Seeing Jesus in Toast: Neural and behavioural correlates of face pareidolia, *Cortex*, Vol.53, pp.60-77.

Mandik, P. (2017) 'Meta-Illusionism and Qualia Quietism', in Frankish, K. (ed.) *Illusionism as a Theory of Consciousness*, pp.140-148. Imprint Academic: Exeter.

Marr, D. (2010) *Vision: a computational investigation into the human representation and processing of visual information*. MIT Press.

Marshall, D. and Weatherson, B. (2018) 'Intrinsic vs. Extrinsic Properties', in Zalta, E. (ed.), *The Stanford Encyclopaedia of Philosophy (Spring 2018 Edition)*, Available at: https://plato.stanford.edu/entries/intrinsic-extrinsic/ accessed 30/07/2017.

Mazur, J. (1983) 'Response Optimisation: A Result or a Mechanism?', *Science*, 221(4614), p.977.

McGurk, H. and MacDonald, J. (1976) 'Hearing Lips and Seeing Voices', *Nature*, Vol.264, pp.746-748.

Kahneman, D. (2012) *Thinking Fast, Thinking Slow*. Penguin.

Nagel, T. (1974) 'What Is It Like to Be a Bat?', *The Philosophical Review*, Vol. 83, No. 4, pp.435-450.

Niikawa, T. (2021) 'Illusionism and Definitions of Phenomenal Consciousness', *Philosophical Studies*, 1, pp.1-21.

Noë, A. (2002) 'Is the Visual World a Grand Illusion?', in Noë, A. (ed.) *Is the Visual World a Grand Illusion?* pp.1-12. Imprint Academic: Exeter.

Pereboom, D. (2017) 'Illusionism and Anti-Functionalism about Phenomenal Consciousness', in Frankish, K. (ed.) *Illusionism as a Theory of Consciousness*, pp.11-39. Imprint Academic: Exeter.

Phillips, W. (2011) 'Cross-Cultural Differences in Visual Perception of Colour, Illusions, Depth, and Pictures', in Keith, K. (ed.) *Cross Cultural Psychology*, pp.160-180, Wiley-Blackwell.

Place, U. (1956) 'Is Consciousness a Brain Process?', *British Journal of Psychology*, 47(10), pp.44-50.

Prakash, C., Stephens, K., Hoffman, D., Singh, M. and Fields, C. (2019) 'Fitness Beats Truth in the Evolution of Perception', *Acta Biotheoretica*, 69, pp.319-341.

Prinz, J. (2017) 'Against Illusionism', in Frankish, K. (ed.) *Illusionism as a Theory of Consciousness*, pp.183-196. Imprint Academic: Exeter.

Ridder, D., Adhia, D. and Vanneste, S. (2021) 'The Anatomy of Pain and Suffering in the Brain and its Clinical Implications', *Neuroscience & Biobehavioural Reviews*, Vol.130, pp.125-146.

Robbins, P.(2017) 'Modularity of Mind', in Zalta, E. (ed.) *The Stanford Encyclopaedia of Philosophy* (Winter 2017 Edition), Available at: https://plato.stanford.edu/entries/modularity-mind/ accessed 15/01/2023.

Robinson, Z, Maley, C., and Piccinini, G. (2015) 'Is Consciousness a Spandrel?', *Journal of the American Philosophical Association*, 1(2), pp.365-383.

Ross, A. (2016) 'Illusionism and the Epistemological Problems Facing Phenomenal Realism', *Journal of Consciousness Studies*, Vol.23, No.11-12, pp.215-223.

Sahu, M. (2022) 'Representationalism, Scepticism, and Phenomenal Realism: an appraisal of the non-reducibility of phenomenality', *Prometeica*, 25, pp.51-65.

Schleidt, W. and Shalter, M. (2003) 'Co-Evolution of Humans and Canids: an alternative view of dog domestication: homo homini lupus?', *Evolution and Cognition*, Vol9., No.1, pp.57-72.

Schwitzgebel, E. (2017) 'Phenomenal Consciousness Defined and Defended as Innocently as I Can Manage', in Frankish, K. (ed.) *Illusionism as a Theory of Consciousness*, pp 224-235. Imprint Academic: Exeter.

Sellars, W. (1962) 'Philosophy and the Scientific Image of Man', in Colodny, R. (ed.) *Frontiers of Science and Philosophy*, pp.35-78. University of Pittsburgh Press.

Shabasson, D. (2021) 'Illusionism about Phenomenal Consciousness: explaining the illusion', *Review of Philosophy and Psychology*, 13, pp.427-453.

Shelton, J., and Kumar, G. (2010) 'Comparison between Auditory and Visual Simple Reaction Times', *Neuroscience and Medicine,* Vol. 1, pp.30-32.

Strawson, G. (2018) 'The Consciousness Deniers', *The New Yok Review of Books*, available at: http://www.nybooks.com/daily/2018/03/13/the-consciousness-deniers/ accessed 15/06/2018.

Tye, M. (2021) 'Qualia', in Zalta, E. (ed.) *The Stanford Encyclopaedia of Philosophy* (Fall 2021 Edition), Available at: https://plato.stanford.edu/entries/qualia/ accessed 02/12/2024.

Tversky, A. and Kahneman, D. (1974) 'Judgement Under Uncertainty: Heuristics and Biases', *Science*, 27, 185, pp.1124-1131.

Williamson, T. (2020) *Philosophical Method*. Oxford University Press.

Wittgenstein, L. (2009) *Philosophical Investigations*, 4th Edition. Wiley-Blackwell.

Zimmerman, E. and Lange, J. (2022) 'Saccade Suppression of Displacements, but not of Contrast, Depends on Context', *Journal of Vision*, Vol. 22, No. 10.

WHERE IS MY MIND?

In this paper I argue in favour of radical cognitive extension, sometimes referred to as cognitive bloat. According to the Extended Mind Thesis (Clark and Chalmers 1998), under the right conditions cognition and the mind can extend over external artefacts like notebooks. Various critics—along with Clark and Chalmers 1998—seek to limit the scope of the thesis, often appealing to a reductio ad absurdum. However, rather than the implications of the Extended Mind Thesis being absurd, I claim they are counter-intuitive but revealing insights into the nature of cognition, and processes in general. In this paper I argue against Clark and Chalmers' (1998) 'Availability and Portability' Criteria and consider a parallel case of Extended Digestion and the extent to which we consider entities apt for a particular functional role is based on human-centric concerns, to make the case for a wide view of cognitive extension, in favour of cognitive bloat.

1 Introduction

"There are many, many ways to carve the world... Think of the way we identify different men with their bodies, with the position they occupy in place and time. Since we inherit this way of thinking, we assume that it is natural, that it is the only way. But what if we identify a man with his thoughts—what then? How would we draw his boundaries? Where would he begin, and where would he end?"

—R. Scott Bakker, *The Judging Eye*

Where can you find a mind? Where is my mind? A typical answer might be: a mind is what the brain does, and therefore minds can be found inside our heads. Not everyone agrees with this answer. Andy Clark and David Chalmers (1998) argue that minds are not just inside heads, that they can in part reside in the external world. They claim

minds—including yours and mine—can extend across everyday objects in the environment, things as mundane as a humble notebook.

Clark and Chalmers' Extended Mind Thesis rests on the idea that under the right conditions an object, such as a notebook, can be used as part of a cognitive process that would otherwise be carried out solely by the brain. This would make the object not just a tool or source of information, but part of ongoing cognitive processes, and so (as Clark and Chalmers go on to argue) part of our minds.

Understanding where the borders of our minds are is important. For example, if objects in the external world can be a part of our minds, it raises a range of issues around ownership and access to those external objects. Directly interfering with someone else's brain can be bad, because it could directly affect their mind, affect *them*. If an evil genius interfered with *your* brain to change *your* decisions, to tamper with *your* memory, it would be a very bad thing indeed. If the Extended Mind thesis is right, and external objects do form part of our mind, perhaps it is also a very bad thing to interfere with such external mind-objects. This raises the prospect that you may have rights relating to, and other people may have obligations towards, these external aspects of your mind.

If we accept that objects like notebooks can be used as part of our cognition, it may be the case that *other brains* can carry out part of 'our' cognition as well. Under the right conditions, part of the answer to "Where is my Mind?" may be: "inside other people's heads". Perhaps even in *your* head. If our minds really can extend into other people's heads in this way, it might lead to difficult questions of cognitive ownership and access.

I argue that the claim that you can have a part of my mind in your head, or at least part of my cognition in your head, is correct. In this essay I evaluate Clark and Chalmers' (1998) Extended Mind Thesis and advance the claim that you may have a piece of my mind. As part of this I will be arguing that the typical views of cognition as being brain-bound are overly conservative.

1.1 In Favour of Cognitive Bloat

There are various criticisms of the Extended Mind Thesis. Some of the criticisms seek to deny that interactions with external objects can count as cognition (e.g. Adams and Aizawa 2001 and 2010). Other criticisms point to apparently absurd consequences if the Extended Mind Thesis were true, like the Internet being able to carry out parts of our cognition (Spevrak 2020, p.6). I think we should bite the bullet of these apparently absurd consequences of the Extended Mind Thesis, of what is sometimes called 'cognitive bloat', and accept them as true.

My view rests on the core theory of the Extended Mind, articulated in the Parity Principle, that it is possible to substitute parts of a cognitive process, providing the substitution can carry out (more or less) the same functional role as the substituted part. In arguing for my view, I reject limitations placed on cases of cognitive (and mind) extension by Clark and Chalmers (1998 pp.9-12, and Clark 2010, p.46) and argue in favour of cognitive bloat. Typically, cognitive bloat is seen as an absurd and unwelcome implication of the Extended Mind Thesis. It runs counter to our intuitions of how we normally think about how things work. Instead, I claim it should be seen as illuminating as to the true picture of cognition and the role that external entities can play in it.

It is natural to think of cognition as residing in our heads, bound by the borders of skin and bone. We are inclined to think of the world in terms of objects, and not in other terms such as flows of information and energy. When we divide up the world into various kinds, including the cognitive and non-cognitive, we are inclined to draw our dividing lines at the edges of physical objects. While this is useful, and what we have in our heads and only in our heads is significant and important, it does run the risk of overlooking what the reality is. Cognition, and other bodily processes, outrun the borders of our bodies, and how this occurs and the impact it has on us deserves attention.

I make two positive arguments in favour of cognitive bloat. The first argument is that the Extended Mind Thesis is not making some unusual or exotic claim, rather it is bringing to our attention just one example of how processes extend beyond the physical borders of an object. To illustrate this, I use the analogy of Extended Digestion, to make a case for accepting extension of bodily processes across external objects. By focusing on a case that is less philosophically loaded than cognition, I hope to make mind extension seem less absurd. If one accepts the Extended Digestion Thesis, it makes the Extended Mind Thesis harder to reject. If bodily process extension is just a natural phenomenon, then critics of the Extended Mind Thesis must offer principled reasons for cognitive exceptionalism.

The second argument is intended to make the case that cognition is more commonplace than we typically think and not entirely restricted to brains, and so we should not be so robust in our gatekeeping as to what counts as cognition and what

does not. What matters is whether cognitive phenomena are occurring, not what human-centric considerations and perspectives are at play. The case for this position is made via an argument by analogy considering doorstops, and how various objects can be used as a doorstop and the context-sensitive criteria under which we recognise objects as doorstops. While doorstops are not cognitive, the analogy aims to highlight how some of our classification schemes—like whether something counts as cognition—are grounded in how things seem to us and our human-centric concerns, rather than principled criteria about the phenomenon.

A reasonable worry about my view—and about cognitive bloat in general—is that it is too liberal a view and leads to too much being counted as cognition. I think this worry is misplaced, and instead we should accept that there is much cognition or potential for cognitive extension out there in the world. Once we do this, I think the more interesting question is "whose cognition is it?". I provide a tentative approach to answer this question, based on the Parity Principle, and whether the external part of the process is being driven, or carried out, *by* the relevant agent or not.

As a consequence of the view set out here, I have a much broader answer to the question of 'Where is my Mind?' than that offered by Clark and Chalmers (1998). This broader answer includes a range of external artefacts, which might be rejected by those who view cognitive bloat as an unacceptable implication of cognitive extension. One consequence of my view is that if you are willing to remember for me that "cheese does not make for a good doorstop", then I will have given you a piece of my mind.

2 The Extended Mind Thesis, Presented

While it is uncontroversial that we use external objects as aids to our thinking, the Extended Mind Thesis goes a step further. It claims that in certain cases the external objects are not just being used by our thinking, they are *part* of our thinking. As a result, "cognitive processes ain't (all) in the head" (Clark and Chalmers 1998, p.8). This is a radical idea. It entails that components of thoughts, as different parts of cognitive processes, are distributed across our brain and external objects.

Below, before formally introducing the Extended Mind Thesis, I outline just how radical the thesis is. Because it is a radical thesis about cognition and mind, it requires us to consider cognition in a way that allows us to properly assess whether something counts as cognitive. As the Extended Mind Thesis is directly challenging our ordinary concepts of cognition, it won't do for a critic to say that external objects like notebooks cannot be cognitive because they are not like familiar cognitive things (brains) as that is the very point the thesis is contesting.

2.1 The Extended Mind Thesis is Radical

The Extended Mind Thesis is radical in two ways. The first I have already touched upon: the idea that our cognitive processes do not take place just in our brains, nor even just within our bodies. According to the thesis, at times they are partly constituted by things in the external world, beyond the borders of skin and bone. The idea that our cognition is in some sense spread outside of us and resides (in part) in the objects we use is strange and counter to what most people would consider common sense. The

Extended Mind Thesis paints a picture of a world in which cognition is all around us, not just contained within discrete entities like people.

The second way in which the Extended Mind Thesis is radical is how it claims cognitive and mind extension occurs. Philosophy and science-fiction is filled with stories and thought experiments about how human minds may extend or reside outside of our fleshy organic brains. To name a few we have: tales of brains removed from bodies yet controlling them from afar via radio waves (Dennett 2009); neurons replaced by machine equivalents (Chalmers 2010); computer chips connected directly to the brain (Clark and Chalmers 1998); minds stored in chips or broadcast across the galaxy to be 're-sleeved' in new bodies or reincarnated into new forms (Zelanzy 2004 [1967], Morgan 2002); people absorbed—borged—into hive minds (Star Trek 1989); and minds uploaded to computer-simulated afterlives (Banks 2010). What these cases have in common is some impressive achievement of science or invasive medical procedures to make the mind extension (or transplant) happen. In contrast, according to the Extended Mind Thesis, mind extension is a mundane and ordinary occurrence. Our brains and bodies already have everything they need for minds to extend into the external environment. Nothing particularly special, not even learning the right techniques, is required for mind extension to occur. I do it, you do it, everyone does it. All it takes is the right kind of interaction with an external thing that can contribute to a cognitive process. Rather than radio-waves or chips in our heads, the bridge between brain and external artefact is our regular sensorimotor interaction with the world: perceiving and acting. Behaviours as everyday as reading and writing lead to cognitive extension.

According to the Extended Mind Thesis cognitive extension is easy. It just sort of happens. And it happens often. Without trying to do it or noticing that it is happening it just occurs. You are likely to have had your cognition extended many times already today. Your mind has been all over the place! Once aware of it, you can notice it occurring, and take a measure of control of it, but generally it just happens without any effort or attention.

2.2 The External World and Cognition

2.2.1 We Use the External World to Aid Our thinking

We all use the world, including our bodies, to aid our cognition. One might say we cannot help but do so. Sometimes we do it deliberately, sometimes we do so unconsciously. We do it so regularly that we do not give it much consideration, but our cognition-dependent behaviours would be severely limited without the use of these external resources.

There are a wide range of everyday interactions with the world that we carry out which aid our thinking. When learning to count we use our fingers as a visual aid. We rearrange jigsaw pieces to make it easier to spot how they might connect with each other. We write philosophical essays not just to communicate our thoughts to others, but also to develop and examine our thoughts. Actions like these are ‘epistemic actions’ intended to improve our cognition (Kirsh and Maglio 1994).

What these examples have in common is that they are not merely a case of an input or an output to our cognition. They are an active engagement with the world, not a

passive sensing of it. We are affecting the world around ourselves to feed the outcomes back into our cognition in order to think better. They are part of a loop of cognitive processing, and without them our cognitive and behavioural performance would be impoverished.

2.2.2 We Make Things to Do Some of Our Thinking for Us

As our technology has advanced so has the use of the external world to aid our thinking, or to use it in place of our thinking. From cave paintings to writing, from word processors to super-computers, and even to how we arrange our workspaces, we co-opt external resources to assist us or do some part of our cognitive tasks. Just as we use machines to offload or enhance physical tasks, so we use machines to offload or enhance cognitive tasks.

Clark and Chalmers (1998, p.8) discuss a useful example of using external resources as part of a cognitive process, via the game Tetris. Tetris is a computer game where simple two-dimensional shapes made of aligned squares fall down a computer display. When shapes reach the bottom of the display, or the top layer of already fallen squares, they halt. At that point, if a complete line across the screen is formed, those squares that form the line disappear. This is the objective of the game: to make as many lines disappear as possible before the shapes run out of room to fall, which brings the game to an end. The element of play is to rotate and move the falling shapes, such that they best plug into the existing strata of fallen squares, to remove further lines.

A key task in Tetris is for the player to work out where falling shapes will best fit. One way of doing this is to mentally rotate the shapes, to imagine what they will look like in different orientations with respect to the rugged terrain of squares at the bottom of the screen. Tetris provides support for this task; with a tap of a button players can make the shapes rotate on the display. This is achieved much faster than rotating the shape mentally (Kirsh and Maglio 1992, p.1). Thus, players can play better—faster—by offloading the mental rotation task to the computer. As well as being faster it takes less mental effort for the player. Setting aside any interaction errors (e.g. accidently pushing the wrong button), getting the computer to do it is also more accurate as its rotations are error free.

2.3 The Extended Mind Thesis

Now that I have sketched out a view of how we routinely use objects in the external environment to support our cognition, I turn to the Extended Mind Thesis, which claims that in the right circumstances these external artifacts constitute part of our cognitive processes. The Extended Mind Thesis has a core theory—the Parity Principle—with its area of application described by the 'sensorimotor liberation statement' on sensorimotor interaction. The thesis also has an auxiliary theory, the Availability and Portability Criteria.

2.3.1 The Thesis

The Parity Principle

We perform a range of cognitive functions with our brains. Many of these functions can in whole or in part be carried out by things that are not our brains. We routinely offload some cognitive tasks, or aspects of them, to our technology and environment, like using a calculator or rotating shapes in Tetris. According to Clark and Chalmers (1998 p.8,11), these are cases of cognitive extension.

Clark and Chalmers' (1998, p.8) core argument for mind extension rests upon mental functionalism, and what has become known as the **Parity Principle**[43]:

> "if, as we confront some task, a part of the world functions as a process which, *were it done in the head*, we would have no hesitation in recognising as part of the cognitive process, then that part of the world *is* (so we claim) part of the cognitive process".

Or, as Clark (2010b, p.93) paraphrases Daniel Dennett: "cognition is as cognition does". That is, if something is doing things we recognise as being just like a cognitive process, then it should be considered cognitive. And one way we can tell when something is doing cognitive things, or is supplying part of a cognitive process, is when it is substituted for part of own cognitive activity that we uncontroversially consider to be cognitive. Thus, if we interact with a calculator or Tetris game in the right way, the

[43] Sprevak (2009, p3) refers to it is the Fair Treatment Principle, emphasising it calls for internal and external cognition to be given equal treatment.

result is that our cognitive processes are split across our brains and devices, bridged by our sensorimotor interaction with the devices' interfaces. According to the Parity Principle, this is not only an extended human-machine system, but also an extended cognitive system. In the case of Tetris, part of the cognitive task that we previously carried out in our heads—mental rotation of shapes—is now being carried out by an external artefact but is still part of the overall cognitive process.

Functionalism

For the Parity Principle to have much traction, it needs to be grounded in functionalism (Wheeler 2010, p.245), so that what matters is what a system is doing, rather than what it is made of. Functionalism about the mind, or metaphysical functionalism[44], is the theory that mental states are functional states determined by the typical causal relationships between their inputs (sensory data and other mental states) and outputs (behavioural and influence on other mental states) (Block 2007, p.28; Knobe and Prinze 2007, p.658; and Levine 2023).

If we accept functionalism, it follows that any state that fulfils the causal role associated with a given mental state is a realisation of that mental state (Putnam 1973). This is the case no matter how that state is implemented, even if it were done with cogs and gears or flows of water, rather than with neurons. For example, the mental state type 'pain' is a function that takes a range of physical insults to the body as its input state, and as its output it typically causes other mental states such that the belief that one is in pain, the desire for it to stop, and typically causes behaviour like

[44] Henceforth I shall just use the term 'functionalism'.

crying out, grimacing, and quickly moving away from the apparent cause of the pain[45]. These input-output mappings to a range of mental states and behaviours are the causal role of pain.

In this example, 'pain' is a mental state *type,* and the various instantiations or realisations of the type are its *tokens* (Block 2007, p.28). Mental state types are multiply realisable; they can be instantiated via different physical means. Thus, men, mice, and Martians may have the same thought, despite having different biological structures that implement them. In Extended Mind cases, the view is that some of our mental activity is being implemented using co-opted external resources (like calculators and notebooks) that participate in realising a mental function, despite being external to us.

Sensorimotor Interaction

Next, we turn to the linkage between internal (brains) and external (calculators, notebooks, and the like) cognitive realisers that enable cognitive extension. The kind of interaction that brings about extended cognition, according to the thesis, is not some fancy technology, but is achieved via regular everyday interactions with the world:

> "A subject's cognitive processes and mental states can be partly constituted by entities that are external to the subject, in virtue of the subject's sensorimotor interaction with these entities" (Chalmers 2019, p.13; and Clark 2019, p.268).

[45] And the various outputs can lead to further state changes.

Sensorimotor interaction is a back-and-forth loop of sensing and acting upon something in the world. We observe, we decide, we act, and then we go back around the loop again as we observe the outcome of our actions. In this way, we are affecting the world to create a new informational state that can be used as an input into our cognitive processes (Clark 2008, p.131). Sensorimotor interactions are the regular way in which we interact with the world, whether catching a ball, typing at a keyboard, or talking to someone. However, it is 'loopy' cases where our actions are changing what gets fed back into cognition that make for sensorimotor interactions that support cognitive extension.

Causal Coupling

Sensorimotor interaction is the means by which brain-bound cognitive resources and processes are *causally coupled* to external resources, creating a "cognitive system in its own right" (Clark and Chalmers 1998, p.8). According to Clark and Chalmers (1998, p.8) a coupled system is one where the components:

1. Have a two-way interaction.
2. Jointly govern behaviour (if the external component is removed, the system's behavioural competence will drop).

The motivation here is to pick out those cases whereby the external resources are being used in such a way that they constitute part of a cognitive process, rather than being merely used by or interacted with by that system. The idea seems to be that by contributing to the activity of the system, those components could be said to be part of it.

For there to be a two-way interaction between two sets of system components, there must be a bidirectional causal flow, which rules out asymmetric influences (Menary 2010, p.3). This two-way interaction must include an active causal role, which has "a direct impact on the organism and its behaviour" (Clark and Chalmers 1998, p.9). Being two-way, the interaction must involve inputs and outputs. This links to the second criterion; the coupling between the components must be productive in some way. In the case of mind or cognitive extension, the coupling must contribute to some cognitive or mind-related behavioural outcome such as making better decisions, having an improved memory, or being quicker at solving maths problems.

The general idea here is that via a sensorimotor bridge—our regular ways of interacting with the world—an external element is coupled with our internal cognitive core. When this happens, the result is an extended cognitive system. What makes the two elements joint members of a cognitive system, rather than one being merely an input or output, is the two-way interaction, and joint contribution to the system's competence for cognitive function. Together they form a new functional arrangement, a new way of carrying out the relevant cognitive process. Thus, merely reading a road sign is an input (a 'flow' in systems-speak) into the cognitive system, whereas using a calculator may be a case of cognitive extension because it involves a two-way interaction that jointly governs behaviour.

2.3.2 From Cognition to Mind

The Parity Principle is used to argue for extended cognition, but what about extended minds? In one sense, in making a case for extended cognition a case for mind extension

has already been made. This may not be enough though. Cognitive processes are often associated with the kinds of things that computers can do, such as rotating shapes in the game Tetris. When we think of cognitive processes in terms of the Computational Theory of Mind, we tend to think of mental activity that can be decomposed into discrete chunks, broken down into a mechanistic algorithm. Perhaps these algorithmic mental activities can be extended into the external world, but not everyone might accept these as a case of mind extension. And so, the challenge may still linger; the Parity Principle is all well and good for extending sub-personal 'mechanical' cognitive processes—like the kind of arithmetic we would use a calculator for—but what about extended *minds*? The common-sense notion of the mind includes things like *thinking*, *believing, hoping*, *loving*, and *hating*. What does the Extended Mind Thesis have to say about these person-level processes and states?

The Case of Otto and his Notebook

Clark and Chalmers (1998, pp.12-14) respond to this challenge by giving an example of extended belief. They describe a thought experiment about a character by the name of Otto. Otto has Alzheimer's, which affects his memory. Because Otto can no longer rely on his memory, he uses a notebook as an external store for information that he wants to have readily available. Key information that he wants to retain goes onto the pages of his notebook rather than—as it used to—gets stored in his brain. Directions, addresses, important dates, passwords galore, and so on, get stored in his notebook. Before Otto had Alzheimer's when he wanted to know something, he would do the same as most people: he would access his memories, get the appropriate information,

and then make use of it. Now, when Otto wants to know something, he accesses his notebook, gets the relevant information, and then makes use of it. The notebook and its contents are as causally relevant to Otto's behaviour as 'normal' memories are to someone else.

Clark and Chalmers (1998, p.13) then introduce beliefs into the picture; Otto acts on the information in his notebook because he *believes* it is true (Clark and Chalmers 1998, p.13). Otto is not just accessing information available in the external world, he has already endorsed the content as being something which he accepts; the notebook is not merely an information-store, it stores the content of his beliefs. When he accesses the information in his notebook, he does so on the assumption that what he reads is true.

Key to this picture is the role which the information in Otto's notebook plays. If we consider the content of the notebook to contain the content of Otto's beliefs, it gives us predictive and explanatory power. We can read what is in the notebook, and if we treat the content as though they were his beliefs, we are able to predict how Otto will behave. We will observe him acting upon true and false statements in his notebooks as though they were his true or false beliefs. We can even change and predict how Otto will behave in the future by changing the content of the notebook.

Further, if we say with his notebook Otto behaves as though he has a certain set of beliefs which correspond to the content of the notebook, when we take the notebook away, we can say that his subsequent behaviour (responses to questions, etc.) corresponds to what we might expect if his beliefs were taken away from him. At this

relatively coarse-grained level of analysis, we can see that the notebook is being used as a kind of external memory store and Otto acts upon its contents as though they were his beliefs. We could introduce extra terms into our analysis (Clark and Chalmers 1998, p.13), such that we list a set of mind states that setup how and why Otto is using his notebook each time, but this adds terms without explanatory gain (Clark 2010, p.46-47).

Some might object that having information in a notebook is not the same as having a belief, that the notebook and its content and how both are used differ in important ways to how we normally conceive of having beliefs. Clark and Chalmers' (1998, p.14) response to this is whether or not the picture of Otto using his notebook fits with standard notions of belief, we ought to accept that Otto has the relevant beliefs. What matters is the role the content in the notebook plays (Clark and Chalmers 1998, p.14). We make our way to a certain street when we want to go to the theatre *because* we believe the theatre is on that street, whether the belief comes from inside or (as with Otto) from a notebook. We wish our friends a happy birthday on the days that we believe to be their birthdays, whether that is because we have memorised their birthdates or because it is recorded in some external resource that we trust to be true. When Otto endorses the content of his notebook, he acts on the content as we would act on our beliefs. Otto's behaviour only make sense if we think he is an agent acting upon his beliefs (Clark and Chalmers 1998, p.13). Given this, as beliefs are mental states, at least some personal-level aspects of our minds can be extended into the external world. The result is that our minds are not just in our heads, they are out there as well.

Coarse Functional Equivalence

A worry is that coupled cognitive systems, such as Otto and his notebook, fail to appropriately realise the relevant mental functions. What is in Otto's notebook, the objection goes, cannot be a memory, or even constitute part of cognitive processes involving memory, because it lacks the right input-output mappings between perception, mental states, and behaviour. For example, the content of the notebook is not interlinked in such a way that updating one part would lead to revisions of other parts, as we might expect from our biological memory. Similarly, our access to our memory is sub-personal. Unlike if we were to use a notebook, we have no awareness or involvement in *how* our regular memories are retrieved. The view here is that whatever Otto is doing with his notebook, it cannot involve the notebook constituting parts of cognitive processes involving memories and beliefs.

To address this concern, we can adopt a coarser-grained functionalist approach (Sprevak 2009, pp.5-11). On a coarser view while mental states are still defined by their causal role, there does not need to be a fine-grained equivalence to realise the role, rather we need to pick out what is essential to a particular mental state type (Sprevak 2009, p.5-7). Hilary Putnam (1965, p.29) gives us an example of Super Spartans, who can supress all outwards signs that they are experiencing pain. The pain state of a Super Spartan is quite different from my own, yet we recognise that they do experience pain and have the mental state of pain. Thus, mental states can vary considerably providing they fulfil their essential roles. Similarly, if we are too narrow in what we accept as a functional substitution we might have to deny that Martians (with similar

but different psychology to ours) could have memories or beliefs, because none of the roles of their mental states are close enough to our own (Clark 2010, pp.44-45 and Sprevak 2009 pp.5-11).

Adopting a coarser-grained functionalist approach allow us to consider mental states at a level general enough to accommodate sufficient variability, such that we can accept humans, Martians, and Super Spartans as being able to have (broadly) equivalent mental states. This generality fits with the extended mind thesis, which does not require extended processes to have the same functional profile as the unextended equivalents (Bayne 2022, p.128). What should matter is whether what is in common between the two is essential to their identity as a particular mental state or process (Bayne 2022, p.128). In the case of Otto, we have recognisable patterns of cognition-dependent behaviour that if it were not for our interest in his notebook, we would happily apply terms such as ‘memory’ and ‘belief’. This indicates that there is sufficient commonality to claim broad functional equivalence, until shown to be otherwise.

2.4 Supersizing and Restraining the Mind

Clark and Chalmers (1998) have made a strong case for the Extended Mind Theory. But how far does the mind extend? How much extension is in the world around us? These are questions to which I will develop answers in the rest of this paper, but first in this section I cover Clark and Chalmers’ response to these questions.

Clark and Chalmers (1998) take a liberal approach to extension. In their original paper, they suggest that computer files, the internet, and other thinkers could be part of our extended mind. I agree. Their paper effectively argues that if standing beliefs or

nonconscious cognitive processes are part of the mind, then the mind can be indefinitely extended to notebooks, external computing devices, and even parts of others' minds (Gertler 2007, p.202 and Clark 2008, p.161). But not everything in the world is cognitive, and so if we are to accept that cognition and minds extend beyond the borders of skin and bone, we need to know where the real borders are, where the extension runs out. We need to distinguish between what interactions with the external world result in extension and which do not. We need to be able to distinguish between the cognitive and non-cognitive.

2.4.1 Availability and Portability Criteria

To set boundary conditions on what external resources can or cannot couple with and extend our 'core cognition' ("the true cognitive processes … that lie at the constant core of the system; anything else is an add-on extra") Clark and Chalmers (1998) introduced the Availability and Portability Criteria. Later, these were updated by Clark (2010, p.46) as a set of 'rough and ready' criteria, which would lead to a 'modestly intuitive' set of candidate cognitive extensions:

#1 That the resource be reliably available and typically invoked. (Otto always carries the notebook and won't answer that he "doesn't know" until after he has consulted it);

#2 That any information thus retrieved be more or less automatically endorsed. It should not usually be subject to critical scrutiny (unlike the opinions of other

people, for example). It should be deemed as trust-worthy as something retrieved clearly from biological memory; and

#3 That information contained in the resource should be easily accessible as and when required.

What Clark and Chalmers are targeting for mind extension with these criteria is a causally coupled system that has a package of cognitive resources that can be brought to bear on a range of problems. Their position is that external cognitive resources, like Otto's notebook, should be just like brain-bound cognitive resources in that they are (nearly) always readily available (#1) and accessible (#3)[46] for use. As a result, when the situation warrants it, the external resource is used, just as a brain-bound cognitive resource would be (#1).

While we should not lose sight of the fact that these are given as 'rough and ready' criteria, we can see the point of them is to capture candidate extension cases whereby the experience or high-level nature of the interaction is akin to using our natural brain-bound resources. Both Clark (e.g. 2008) and Chalmers (2019) talk of paradigmatic cases of extension, such as Otto and his notebook, in terms of there being an automatic sub-personal interaction with the external object, where the interaction is *fluent* (Clark

[46] I take the distinction between Available (#1) and Accessible (#3) to be as follows: availability refers to the presence of a resource and its potential to be used, and accessibility is the ease of getting an output from it. External information may be available but not readily accessible, or in a highly accessible state but not available to me. When I try to recall someone's name, and almost have it but cannot quite get it, when it is on the 'tip of my tongue', the name is available to me, but not currently accessible (Foster 2009, p.26).

2019, p.268), as though the external resource is being used and relied upon as 'just another' resource in the cognitive toolkit.

The requirement for the information involved to be endorsed and trusted (#2) is intended to make it part of the cognitive activity (like biological memory), rather than an external information input for the cognitive system. It is not new 'foreign' information to the system, to be evaluated and potentially rejected, instead it is given equivalent standing as information already stored in our brains, something already within our circle of trust, the borders of what we consider to be 'ours'. Perhaps this is meant to be like making use of a recalled fact, rather than reading something in the news and pausing to consider its credibility. Clark and Chalmers are aiming for an interaction such that it is "almost sub personal in nature" (Clark 2019, p.268), and "…more akin to information flow within the brain..." (Clark and Chalmers 1998, p.16) than merely accessing an external information source.

The purpose of the Availability and Portability Criteria seems to be to (i) narrow the range of possible mind-extensions to those cases that more closely align with what philosophers (and cognitive scientists) would commonly accept as cognition, while also (ii) avoiding the inclusion of brief transitory couplings. These boundary conditions limit the range of cognitive extensions. They attempt to limit what is sometimes called 'cognitive bloat', the ascribing of cognition to so much that the worry becomes that the concept of cognition bloats beyond usefulness. Even with these constraints, some critics think that cognitive extension should be denied or limited further, as will be explored in the following section.

3 The Extended Mind, Defended

"My colleague Ned Block likes to say that the thesis was false in 1995, when we wrote the article, but it has since become true with the advent of smartphones and the like".

— David Chalmers, *Clark and His Critics*

We have met the Extended Mind Thesis. It seems counterintuitive and in opposition to our ordinary assumptions about the boundaries of the mind, even before considering its philosophical implications. Given its radical nature, it is unsurprising that a range of objections to the Extended Mind Thesis have been raised. In this section I respond to some of the prominent objections.

The objections that I consider fall into three broad categories. The first attacks the notion that cognition could spread via a coupling relationship. This objection is about what constitutes a system, and the idea that something could in some sense become cognitive based on how it is used or what it interacts with.

The second type of objection questions whether external processes and vehicles of that process can be considered cognitive if the process is not being carried out in a recognisably cognitive way. In part this objection is about how functionalism is applied to cognition, and how true to the original a replacement needs to be (or how tightly specified the functional role of the mental process needs to be).

The third group of objections rejects the implications of the Extended Mind Thesis. Typically, they point out that if we accept cases like Otto and his notebook, then we would need to accept cognitive bloat, which would lead to a dramatic increase in the scope of external cognition (and mind). Objections of this sort point to consequences

of the Extended Mind Thesis that seem absurd and unacceptable, such that they motivate a rejection of the thesis due to the sheer expansion of what would count as part of our cognition; an *incremento ad absurdum*. My overarching response to this line of objection is to accept the apparently absurd consequences and start to build the case *for* accepting cognitive bloat.

3.1 The Coupling-Constitution Fallacy

3.1.1 The Objection

Adams and Aizawa (2010) argue that Clark and Chalmers' (1998) claimed examples of cognitive extension commit what they call the coupling-constitution fallacy. They hold that cognitive extension could well be possible in theory, but that Clark and Chalmers' examples, such as that of Otto and his notebook, are not actual instances of cognitive extension, and that thinking so is an error brought about by fallacious thinking.

The claimed fallacy is the unwarranted move from there being a coupling between two entities to claiming that one of the entities is part of the other. If X is coupled to Y, say Adams and Aizawa (2010, p68), it does not make X a constitutive part of Y. The point they seek to make is that you cannot make a larger cognitive system (X and Y), by adding a non-cognitive part (X) to a cognitive part (Y) (Adams and Aizawa 2010, p.68). Arguing by analogy, they point out that just because neurons (X) can be coupled to muscles (Y) the neurons do not become part of the muscles or *vice versa* (Adams and Aizawa 2010, p.68). Furthermore, they note that the activity of the neurons that lead to a muscular contraction do not make the neuronal activity a type of muscle contraction (Adams and Aizawa 2010, p.68). There is no 'spread' in the type of activity

taking place. We have neurons doing neuronal things, and muscles doing muscular things, but not neurons doing muscular things nor muscles doing neuronal things.

Simply put, Adams and Aizawa point out that sticking two different things together does not change the nature of those things, and so coupling our brain-bound cognition (X) to something that is non-cognitive (Y) does not impart cognitive status to that something (Y) (Adams and Aizawa 2010, p67). For there to be a larger combined cognitive system the addition (Y) needs to have a cognitive nature (Adams and Aizawa 2010, p.68).

From Adams and Aizawa's perspective, Clark and Chalmers attribute cognitive status too easily, and without sufficient justification. They say that if one accepts the case of Otto being coupled to his notebook as a case of extended cognition, then one should also accept many other couplings as examples of cognitive extension. On this basis we would need to consider a rock cognitive, because one can be coupled with a rock in such a way that it meets the Availability and Portability criteria (Adams and Aizawa 2010, p.68). The implication is that such a seemingly absurd outcome of the extended mind thesis—a rock being part of our cognition—should tell us that there is something wrong with the thesis.

3.1.2 Response

In their objections, Adams and Aizawa are right to bring to our attention that merely coupling two entities together does not change their underlying nature or what they do. They are also right in that coupling may not entail constituting a system of a

particular type. However, I believe Adams and Aizawa are mistaken in not seeing coupling as a constitutive relationship. In those cases where a system of the right type, i.e. a cognitive one, is constituted by the activity of the original system does 'spread' to the new components, such that they participate in the system's activity. And the way we can tell whether this is the case is based on the resulting behaviour and activity of the system, and an appreciation of whether the individual components contribute to that behaviour and activity.

Coupling is a Constitutive Relationship

If we slipped a suitably massive object into an orbital track around our sun, it would become part of the solar system, not just be within it. If we add some carriages to a locomotive engine, we get a train. When we bring people together to work on a joint project, they become a team. Coupling *always* entails some level of constitution (because if nothing else the coupled entities would constitute a set), but just what has been constituted depends on the nature of the things being coupled and how they are coupled. Thus, there is no fallacy in supposing that when two entities are coupled, they form a new (or extended) system, there is only potentially a mistake if a false claim is made about what sort of system has been constituted.

Adams and Aizawa are right that connecting neurons to muscle tissue does not make the muscle tissue part of the neuronal system or *vice versa*, but they do form part of *some* system. In their example the coupling of neurons (X) to muscles (Y) leads to there being a musculo-neuronal system (Z constituted by X and Y) with its own set of behavioural competencies. Such musculo-neuronal systems exist—we all have them—

and are constituted by muscles and neurons. Similarly, coupling external entities, such as a notebook, to a cognitive system such that there is an interaction between them leads to some kind of system. Something new has been constituted by the coupling. The question of interest is if both parts of the newly constituted system are involved in cognition.

What has been constituted leads to cognitive behaviour

For Otto-like cases, what is being claimed is an extension of cognition (and of mental states such as beliefs) across the coupled entities. When looking at such a claimed case of extension we should ask whether the extension produces a characteristic set of *cognitive* behaviours. Clark and Chalmers (1998) have already given us the answer to this question, with the Parity Principle and the example of Otto and his notebook. The addition of the notebook changes Otto's overall cognitive behaviour. Specifically, in Otto's case it *restores* typical cognitive function, and in his use of the notebook he acts as though he has a set of beliefs which he would not otherwise have without the notebook. If the notebook was taken away from Otto, his cognition-dependent behaviours would be seriously and recognisably impaired over a broad range of tasks and situations. In some cases, the notebook is *necessary* for Otto to perform a recognisably cognitive task. Further this necessity is not some kind of precondition, like keeping Otto alive or able to see, it is a necessary component of the cognitive task.

In addition to seeing the reliance that Otto has on his notebook to function at parity with others across a range of cognition-dependent tasks, we can also see that at a certain level of abstraction what is going on with the notebook is cognition-like. We can

predict Otto's behaviour by treating the content of his notebook as his beliefs, including changing what is written in the notebook to change Otto's behaviour. While whatever is going on with the notebook is different to what is going on in the brain, in terms of abstract functions we can see that there is a similarity as well as a broad equivalence in contribution to the wider cognitive system at work in Otto's brain. The notebook is clearly being used to enable characteristically cognitive (and mind-related) behaviours.

In summary, coupling does entail constitution, but what kind of system is constituted, and what kind of processes it carries out, depends on the behavioural competencies of the resulting coupled system. The coupling does not make the separate components part of each other, but it does make them part of the same system. Further, if those parts jointly govern the same process, then they are both participating in the process and so the process (e.g. cognition) is spread across objects.

3.2 But Should We Call This Thing Cognition?

3.2.1 The Objection

There is a remaining worry that even though Otto-like cases seem to involve external entities as part of a cognitive process, that whatever is going on with the external entity it is not cognition. Some critics hold that not just anything coupled in the right way to us can serve as a cognitive extension. Under this view an interaction with a notebook may well serve as a substitute for part of a cognitive process, but it is not cognitive itself. The resulting coupling leads to something cognitive connected to something non-cognitive; even if it *seems* to be doing something cognitive it just isn't.

Critics who hold this view (for example, Adams and Aizawa 2010, and Adams 2019) are looking for a 'mark of the cognitive'; some set of necessary or sufficient features that will be present in all cognitive things, which can be used to distinguish between the cognitive and non-cognitive. Adams and Aizawa (2010; and Adams 2019) argue that Clark and Chalmers are not able to claim some external coupled thing is cognitive, or underpins cognition, if they cannot even tell us what counts as cognitive.

On this view, the external entities are only causally related to the cognitive process but not cognitive themselves (i.e. there is coupling but not constitution of an extended cognitive process). What matters is the nature of the things being brought together, rather than what they are doing. Under this view, we are looking for a set of necessary criteria for potential cognition-hood. Therefore, Otto's notebook would merely be a useful and appropriately formatted external information resource that Otto accesses, just as he accesses any other information source. Sure, it is useful. Sure, it fits neatly as a sort of replacement for his memory, but that does not make it cognitive. No more than seeing the moon or the stars makes them part of our cognition.

Distinct Causal Processes

Adams and Aizawa (2010, p.75) claim that in Otto-like cases there are no high or low-level scientific regularities in information processing that cross the boundary between the brain and the external world. They draw attention to instances where we observe a set of phenomena and mistakenly think they belong to the same natural kind, later discovering distinct causal processes are at play leading to similar phenomena (Adams and Aizawa 2001, pp.51-52). For example, we may have broadly similar symptoms of

illness caused by different families of virus. Or we may have heat, but discover the heat is generated in different ways, such as by fire, decomposition, or friction, which we would likely consider to be of different kinds (Adams and Aizawa 2001, p.51). Adams and Aizawa think that even if candidate cases of extended cognition put forward by Clark and Chalmers appear to be involved in cognition, the involvement of external objects is of a different kind to the cognitive processing carried out by a human brain. Adams and Aizawa (2001) think that as the differences between how a computer and the brain carry out activities like playing chess are so substantially different, they cannot be grouped together as a cognitive kind. On this view, considering cases like Otto using his notebook to involve the notebook as constitutive of cognition is a mistake of the same class as thinking decomposition and friction are of the same kind because they generate heat.

The examples Adams and Aizawa give to persuade us that artificial information processing systems are not cognitive speak to the different and incompatible ways in which the brain and external information processing systems encode and process information. For example, they say "the differences in information-processing capacities between the brain and a DVD or CD[47] player is part of the story of why you can't play a DVD or CD with just a human brain" (Adams and Aizawa 2010, p75). These incompatibilities and differences are why in comparison to technological information processing systems "the brain is capable of linguistic processing, whereas these other devices are not", and "the brain is capable of facial recognition over a range of

[47] DVDs and CDs are optical media in the form of flat discs used to store information, typically movies (DVDs) and music (CDs).

environmental conditions, whereas these other devices are not" (Adams and Aizawa 2010, p75).

3.2.2 Response

Causal Processes and Resulting Phenomena

The challenge of the same or similar phenomena being brought about by different causal process raises interesting questions about how we come to know what different things are. But it misses the mark. Adams and Aizawa's (2001, p.51) challenge is roughly "how can you consider these things that generate heat comparable? Decomposition, friction, fire… these are entirely different things", as a comparison to cognition. I agree with their challenge, but cognition is the heat in this example. If the phenomenon of interest is heat, i.e. something like an increase in local temperature, then the underlying causal processes (friction, fire...) may be radically different from each other, but they are similar enough in that they lead to heat. While there are various secondary considerations why it might matter what caused the heat, if all we are interested in is the heat, then what caused the heat is not relevant.

Adams and Aizawa (2001, p.52) think what should matter as to whether something counts as cognition depends on whether the relevant distinct underlying causal processes are in play. Such an approach allows for a distinct field of scientific study with its own theory for a particular natural kind (e.g. we should have a theory of decomposition, a theory of friction, a theory of fire, but not a theory of heat) (Adams and Aizawa 2001, p.51). It should be noted that the advance of technology has not been kind to Adams and Aizawa's examples made in support of their argument. For

example, “facial recognition over a range of environmental conditions” (Adams and Aizawa 2010, p75), mentioned as a stalwart of brain-bound cognitive processing, is now a common-place task for computers. Cameras and smartphones typically come with this capability built-in, not only doing well at face recognition, but face identification as well. If such a task was supposed to be distinctly cognitive, then we have external devices doing distinctly cognitive things.

Adams and Aizawa’s position (while important) is overly focused on the implementation of cognitive processes. What I am arguing for, and what Clark and Chalmers (1998, p.14-15) are arguing for, is that the phenomenon of interest is how and why people behave in a certain way, and how that relates to their thoughts. In cases like that of Otto he has cognition-dependent behaviours that are reliant on and influenced by the content in his notebook, and we better understand and predict such behaviour if we consider the notebook within the boundary of his cognitive system. What matters more here is what something does, in terms of its functional integration and causal contribution to intelligent behaviour (Wheeler 2010, pp.247,253,267-268), than whether we can point to regularities and similarities in information processing. If we consider notions like belief and memory in this wider way, we are better able to explain Otto’s behaviour (Clark and Chalmers 1998, p.14).

Additionally, if the phenomenon of interest is mental states and behaviour, ultimately the nature of the causal processes are in play is not relevant to the question of whether or not there is cognition. We can agree with Adams and Aizawa that Otto using his notebook does not qualify as cognitive as the term is standardly accepted, but

in some ways so much for the standard view; there is cognition-like behaviour here that if it were done in the head we wouldn't hesitate to consider cognitive, and while we should be interested in whether it is 'truly' cognition, in some ways that's missing the point. This response is developed more fully in Section 7.

3.3 Incremento ad Absurdum

If our cognition can couple with and extend across external entities, where does it stop? There is a worry that the extended mind thesis points to cases of extension that are too easy and too common, too transitory and trivial. If we were to accept the thesis, so the worry goes, our cognition would be spread across our environment.

Sprevak (2020, p.6) suggests accepting too much extension would render our mental concepts "pointless, absurd, or otherwise unfit for purpose". He points to extension to a variety of externals—such as libraries, smartphones, the internet, and friends—as unwanted or unwarranted implications of the Extended Mind Thesis. The problem of these unpalatable outcomes is sometimes referred to as 'Cognitive Bloat' (Sprevak 2020, pp.6-7). Sprevak (2009) argues that cognitive bloat is entailed by functionalism, as functionalism leads to radical cognitive extension, and on that basis, functionalism seems to be false. Sprevak (2009, p.24) goes further saying that those who "stubbornly assert the truth" of radical cognitive extension are being dogmatic in the face of the violation of "so many pre-theoretical intuitions" about our mental states.

In this paper, it is my goal to convince you that radical cognitive extension—cognitive bloat—is plausible, or at least make it seem a somewhat less dogmatic position to hold. In contrast to Sprevak and others, I think that cognitive bloat should be embraced as an

interesting feature of how our cognition works. Here I briefly sample some of the objections to cognitive bloat, which come in the form of *reductios*, giving a brief response, before moving on in later sections to develop a broader case in favour of cognitive bloat.

3.3.1 Conjoined Minds

Notto and Otto

Frank Adams (2019, p.28) gives us a twist to the familiar tale of Otto. In Adams' version, it turns out that Otto is conjoined with his twin Notto. In this tale Otto still has Alzheimer's, inhibiting his ability to lay down new memories, but his twin Notto does not. Instead of a notebook, Otto stores information in Notto, telling him everything that would otherwise have been written in his notebook. Now when Otto wants to retrieve information, like the location of the theatre, he asks Notto. We have the same setup as the case of Otto and his notebook, where some external entity is functioning as an information or belief store, but in this case that store is Notto.

Adams (2019, p.28) considers the case of Otto and Notto to be a clear *reductio* of the Extended Mind Thesis. His view is that the processing that occurs in Notto's brain as he responds to a request for information from Otto is not part of an extended cognitive system including Otto. Notto's response is Notto's cognitive processing. Adams does not give any further argument with regards to Otto and Notto, (not unreasonably) considering the apparently absurd consequence of having to accept that Otto's cognition extends into Notto is sufficient to make his case.

There are two elements to Adams' Otto and Notto objection. The first is the apparent absurdity, the second is the problem of overlapping minds. To the first point, a *reductio absurdum* serves to highlight implications that follow from an argument which we may not wish to accept, but it does not establish that the original argument is wrong. Yes, the Otto and Notto case seems even more unintuitive than the Otto and his notebook case, but as I have been and will continue to argue, these apparently absurd implications are actually interesting features of how cognition and other processes can work. To reject the Otto and Notto case we need a principled reason why their cognitive processes could not extend into each other.

For the second point, Adams (2019, p.29) is denying that processing in Notto's brain could be part of Otto's cognitive processing because it is in Notto's brain. However, this in itself is not a barrier for cognitive extension to occur. For example, Otto's notebook could be a shared resource, with his close friend Clara also using Otto's notebook to store important information which she later endorses and relies upon. This sharing would not preclude Otto and Clara from having cognitive processes extending over the same notebook. It might perhaps have secondary implications, like the notebook being less available to each if they are sharing it, but the notebook and its contents can still perform the same role. In the same way, just because Notto's brain is *his* brain is not a barrier for Otto's cognition to extend to it.

Otto certainly has less *control* when it comes to Notto than when it comes to his notebook. Notto can refuse to store information for Otto or refuse to offer it up when needed. He can deliberately deliver falsehoods or be distracted when Otto needs him

the most. In various ways Otto is reliant on Notto's cooperation for the invoked cognitive process to be appropriately extended to Notto. However, when Notto is cooperating in the right kind of way, we have a distinctively cognitive process distributed across the two, bridged via sensorimotor interaction, to set up interaction loops between them, such that Notto's involvement contributes to Otto's behaviour: extension occurs.

The Philosophy Consultant

Katalin Farkas (2012) asks us to consider the case of Lotte, who studies philosophy and achieves excellent grades. However, it transpires that Lotte has achieved her grades via the help of a philosophy consultant she has hired, and to whom she has full-time access via a radio link. The way this tale runs is that the relationship between Lotte and the philosophy consultant meets all the criteria of the Extended Mind Thesis including the Availability and Portability Criteria; Lotte automatically endorses the content from the consultant, the consultant is reliably available, and so on. The result is that if we accept Otto's notebook as part of his mind, then we are under pressure to accept that the philosophical knowledge in the consultant's brain is part of Lotte's mind. In this case they are both a shared resource, even if they play somewhat different causal roles.

Cases like these are raised as objections or problems for the Extended Mind. But—as I argue—rather than accept them as valid objections they should be embraced as telling us something interesting about our minds. I think we should accept the apparently absurd outcome of Farkas' *reductio*. Under the right circumstances, the Lotte mind is a

mind that spans the brain in Lotte's skull and the philosophy consultant's skull, just as is the case with Otto and Notto.

This conclusion may seem strange or even offend those in the business of marking exams or awarding degrees, but it would just be a case of recognising the relevant competent system. It is not a cheat if the degree is awarded appropriately. Now, it may well be the case that we only want to award degrees to a suitably competent human being with the resources that they can typically be expected to have access to, and on that basis we might deny extended resources housed in other bodies. But we need to ask just what the exam taker be typically expected to have access to. Analogous cases may be of open-book exams, or tests that allow the use of calculators, both examples of external resources being employed in demonstrating competence. In these cases, we are recognising the competence of an individual with those external resources. Indeed, part of the test may be to demonstrate competence at making use of these external resources. In some instances, like with the use of a calculator, the exam-takers' mind might be extended over those resources (as per Clark and Chalmers 1998, p.11). Therefore, it is not extension of mind in and of itself that is a matter of concern for tests and exams, but which externals we consider acceptable and whether they will be generally available and accessible to the individual in future. This is a question of what package of resources and capabilities we want to examine, and we may well rule out a philosophy consultant, perhaps on the basis of him not being generally available or providing too much support such that Lotte may lack a suitable foundation of philosophical ability without him. Whereas textbooks, calculators, and notebooks are fungible items and can be readily sourced and replaced if needed to maintain Lotte's

behavioural competence when it comes to philosophy, the same might not be said of her philosophy consultant.

Overall, I think we should accept cases of cognitive extension to others, like those raised by Adams (2019) and Farkas (2012). I will, however, caveat my position of accepting these cases of conjoined minds somewhat. The Extended Mind thesis makes a claim about cognition at a relatively high level of abstraction, whereby *approximate* external substitutions for components of cognitive processes should be considered as constitutive of our cognition. We should keep in mind that these are approximate substitutions. For example, that Otto endorses the information in his notebook in such a way that it can be considered as a belief does, I believe, make it constitutive of his belief. However, it is a different kind of belief compared to the regular sort, because it only sort-of implements the full functional role of a regular belief. The external objects contribute to the cognitive process in such a way that Otto behaves in a general way as though it is a regular belief, but these extended beliefs are not like our brain-bound beliefs. For example, they are not open to revision in quite the same way as regular belief and do not participate in a relationship with Otto's other beliefs. That is not to say it does not qualify as a belief—we're back to if "it were done in the head, we would have no hesitation in recognising [it] as part of the cognitive process" (Clark and Chalmers 1998)—but we should recognise that it is a kind of sub-species of a regular belief.

3.3.2 The Library

Sprevak (2009, p.16) points out that if we accept the case of cognitive extension between Otto and his notebook, then we should accept that if we step into a library we might acquire millions of new beliefs based on the contents of the books in the library. If Otto can use his notebook in such a way as to extend his cognition, then there seems to be nothing stopping the application of the Parity Principle to the use of a library full of books. Even if we can raise objections to this library-extension case based on what we know about cognition in humans, Sprevak argues that there is a possibility of some alien, e.g. a Martian, being so constituted as to have a memory resource similar to having access to a library of books (Sprevak 2009, p.16). To Sprevak, this and similar cases are enough to reject radical cognition extension (and functionalism).

While I have sympathy with this position, where Sprevak sees a conflict which should be resolved by denying cognitive extension, I think we should go the other way and endorse the radical extension. While the case of a library is more radical than that of Otto and his notebook, we can also see the differences. Otto's notebook resource is relatively small, he has access to it all of the time, he has pre-endorsed its content as true, which is not true of the library. While cognitive processes *could* extend to the library in the same way as Otto's cognition extends to his notebook, as a kind of memory resource the library is much more cumbersome to use. The information is not portable in the way that the notebook is, it takes longer to retrieve relevant information, Otto is less likely to update the content by writing in the books versus his notebook, he may not endorse this third-party content in the same way, and so on. The

sensorimotor interaction with the external resource of the library is much more cumbersome and time consuming, and while the content of the library can contribute to governing Otto's behaviour, it does so in a much more limited way than the content of his notebook. It is closer to being a passive information resource, than a constitutive part of some of Otto's cognitive processes, though that would depend in part on how Otto interacts with the library, the extent to which it would involve loopy interactions.

What this leaves us with is the view that cognitive extension to a library is possible. However, the potential for the library to constitute part of our cognitive processes is likely poorer and more limited than the case of Otto and his notebook.

3.4 Learning to Love Cognitive Bloat

None of the objections covered in this section are fully successful. While they raise useful and interesting points about cognition and systems, they do not successfully undermine the core functionalist argument of the Extended Mind Theory. If we have i) external and brain-bound resources coupled together ii) which jointly enact what we recognise as a cognitive process (perceiving, judging, recalling, etc.), then we have an extended cognitive system. Further, we can have even more confidence that is the case if iii) the external thing involves or contributes to computational processing of representations, such that it iv) contributes to cognition-dependent behaviour. As cognition is part of a mind, and we have Otto-cases that feature more 'minded' content such as beliefs, we have extended minds. It does not matter what those various coupled resources are doing or what they are made of, provided they realise a functional role that is recognisably cognitive. No further mark of the cognitive is

required to establish that cognitive extension can and does occur, though such a mark (if it exists and is definable) would be useful in demarking the boundaries of cognition.

That apparently absurd implications can be drawn from the Extended Mind Thesis cannot by itself prove the thesis false. While they could be part of a set of reasons to reject the thesis, there is (in my view) no strong defeater argument for the thesis. Instead, what these absurd implications do is provoke our intuitions and folk conceptions of cognition and mind, to motivate us to reject the thesis. However, our conception of these things is notoriously flawed and shaped by our cultural and theoretical backgrounds, as well as our subjective perspective. The literature on cognition and the mind is filled with empirical findings in conflict with our folk understandings of our own minds. Rather than undermining the Extended Mind Thesis, these apparently absurd implications are the productive outcome of the theory. They point to how our everyday folk conception of the borders of our cognition and minds is too constrained.

And yet, we seem be able to need to be able to draw the boundaries of cognition somewhere. We need to be able to point and say *this* is cognition but *that* is not. I will get to this topic, but first I will seek to extend the thesis further.

4 The Extended Mind, Unrestrained

I introduced the Extended Mind as a radical thesis, and in the previous section I defended it from a variety of objections and criticisms. In this section I will seek to persuade you of my objection to the standard expression of the Extended Mind thesis. The objection? It needs to be more radical, not less, to properly capture the phenomenon of extended mind. The thesis is unnecessarily restrained by the Availability and Portability Criteria. While I think there is a general case to be made against the criteria, which I discuss in Section 7, to 'soften the ground' for radical extension in this section I make a case that on their own merits the criteria should be discarded.

4.1 The Availability and Portability Criteria

Recall that the Availability and Portability Criteria were introduced (Clark and Chalmers 1998, Clark 2010) to set boundary conditions on what external resources could appropriately couple to and extend our 'core cognition'. The criteria are meant to limit extension cases to those that are close kin to our regular cognition. According to the criteria, the external resources which are extensions candidates should (Clark 2010, p.46):

i. be reliably available and typically invoked;

ii. have content which is more or less automatically endorsed, trusted on par with our brain-bound capacities; and

iii. have the information in the resource be easily accessible as required.

The motivation for these criteria seems to be to avoid watering down the notion of cognition (or mind) to the extent where it loses much of its meaning and practical application. Clark and Chalmers (1998) seem to be seeking to restrict extension to those cases where the information processing taking place is as much alike to our native cognition as can be achieved given the involvement of external vehicles of cognition and the sensorimotor interface between the internal and external components of cognition.

While these criteria may work to mark out less contentious candidates for cognitive extension, they do not integrate well with the core theory based on the parity principle. They pose their own problems and should be discarded. Below I set out why.

4.1.1 Reliably Available and Typically Invoked

Clark (2010) tells us that the external candidate cognitive resource must be reliably available and typically invoked. Clearly, to contribute to the behaviour of a system—whether cognitive or otherwise—a resource needs to be invoked, and to be invoked it needs to be available (it also needs to be accessible (#3) to be invoked successfully). So far, so good. However, we are given no strong reason why the 'reliably' or 'typically' qualifiers are required.

The Substantial Realisation Worry

The reliability and typicality criteria are intended to rule out one-off or occasional interactions. Such interactions could be said to not instantiate a functional property in any substantial sense. The worry here is that if I use some random object to trap a

mouse, or to keep a door open, that does not make the random object a mouse trap or a doorstop. So, while we might want to rule in Otto and his notebook as an acceptable case of cognitive extension, under these criteria we would rule-out Otto's occasional use of his friend Clara's notebook when he cannot find his own. This would be because it is not reliably available (Clara normally has it) or typically invoked (Otto uses his own notebook in preference to Clara's).

In cases like these, despite the worry whether a substantial realisation has occurred, we can say *a* realisation has occurred. We might not go as far as to say the random object put to use as a doorstop is a doorstop, but there is no denying that its effective use leads to a door being stopped[48]. One could object that duration matters, and doubt that it could be possible for something to realise functional equivalence (even at a coarse level) for a short period of time. The thought is that one's mental states and general functional organisation leads one to be disposed to respond in certain ways much further into the future (Block 1978, p.279). There are two general responses to this concern. The first response is to insist that a functionally equivalent substitution has taken place. If we consider the mind in terms of machine table functionalism, (at a sufficiently coarse-grained level) the swapping in and out of different objects like notebooks which are (coarsely) functionally-equivalent does not lead to a change in the machine table, even if a particular object is present for only a brief instance[49]. We have the same configuration of mental states as we had before. The vehicle of cognition has

[48] I return to this example in Section 7.
[49] This somewhat follows Block's (1978, p.279) response to a similar brief realisation worry with regards to his China Brain thought experiment.

changed, but not the causal arrangement to enact the cognitive process. Thus, the causal relationships and dispositions of the mental states (at an appropriate level of coarseness) persist, and hence there is not a problem with the duration of the extension.

The second response is relevant if we consider a stricter requirement for functional equivalence when it comes to cognitive extension. It is to accept that the configuration of mental states (i.e. the machine table) has changed (it is a slightly different machine table, perhaps with states corresponding to a different set of beliefs). At this point in time, the configuration of mental states has dispositions to respond and generate outputs in a particular way, until the configuration of mental states is next changed. In this case, the previous causal relationships and dispositions did not persist, and there is a new set of relationships and dispositions. However, the resulting set of causal relationships and dispositions need not be unstable. It is a new system, so it will behave differently to the previous system, but all things being equal the causal relationships and dispositions of the mental states will persist as they normally would, unless the system is once again reconfigured.

While there may be a more substantial sense of instantiating a functional property than would be the case for fleeting cognitive extensions, this more substantial sense has no bearing on whether *at that point in time* the interaction with the external resource constituted part of the cognitive process. If Otto uses Clara's notebook once, we may have things to say about whether in any substantial sense Clara's notebook is part of

Otto's cognitive resources, but it remains the case that Clara's notebook was part of Otto's extended cognitive process at that point in time.

The Neuron Replacement Teleporter Device

Here is a thought experiment in response to the substantial realisation worry. Suppose one or some set of neurons in my brain were replaced by machines that behaved in the same way as the original neurons. Because they behave in the same way, they can perform the same role. My cognition carries on as before. I can imagine rocks and rainbows and read about rivers and ruins, exactly as I did before my neuron-replacement surgery. The machinery in my brain is reliably available and typically invoked. So far, so good.

Now imagine further that these little neuron-imitation machines themselves need replacing on a frequent basis. They just keep breaking down! So, what happens is that the team of scientists who have been experimenting on me use a teleporter device to constantly remove the machines in my head and replace them with fresh ones. They do this every few seconds. My brain has a constant stream of departing and arriving neuron-imitation machines. As a result, none of the same *token* neuron-imitation machines are readily available or typically invoked. Some of them never even get invoked before being removed and disposed of.

What do we say about this? I suggest what we cannot say is that my cognition is not extended over any of the neuron-imitation machines. Because at certain times some of them are engaged in part of a cognitive process. Further, if they were not there then there would be certain cognitive processes that I could not carry out, or only in a

diminished way. To deny that they constitute part of the cognitive process, when they actively partake in it and are necessary for it, would require denying that cognition is taking place, even if it were indistinguishable from cognition occurring in a normal brain.

The response to this is that the steady stream of neuron-imitation machines teleporting into place in my brain are collectively reliably available and typically invoked, or that they are available and typically invoked as a *type* rather than as a *token*, and so there is a *capacity* that is reliably available and typically invoked, even if not individual elements.

Suppose now that the team of scientists hook up the neuron-machine transporter device to a random number generator. As a result, in an unpredictable fashion there are times when I do have neuron-imitation machines in my brain and I am able to think about rainbows, and there are times when there are not and I cannot. To say that when they are there and are in use their activity cannot constitute part of my cognition because they are not reliably available and typically invoked is to deny that cognition is occurring when I think about rainbows.

Consider also a case where the scientists keep tweaking the neuron-imitation machines such that my beliefs keep changing. At any given time, the beliefs that I would hold would be beliefs in the sense that they perform the role of beliefs in my cognitive economy and have a causal and explanatory role in my belief-dependent behaviour. To deny that I would have beliefs because these belief-candidates were too transitory would leave unexplained what it is enacting the role of beliefs if they are not beliefs.

While we can recognise that under such circumstances my beliefs are not quite like regular beliefs—they are far from being stable—and we might even want to claim they are defective or partial, they're beliefs in that they're enacting the role of beliefs. Similarly, if regular-me with an organic brain has a set of beliefs, we should not deny that some of them count as beliefs if they turn out to be short-lived (perhaps I have a revelation, or suffer a blow to the head damaging some neurons, or an evil scientist interferes with my brain).

4.1.2 Automatically Endorsed

The second of the Availability and Portability criteria is that the information be more or less automatically endorsed when retrieved, deemed as trustworthy as something retrieved from biological memory (Clark 2010). What Clark and Chalmers are aiming for is a kind of sub-personal acceptance of external cognitive outputs, such as the contents of Otto's notebook, as though they were just like our internal cognition. The main issue with this endorsement criterion is that our memories and outputs from our internal cognitive capacities are not always automatically endorsed, yet still count as cognitive, and we should not require a higher standard from external vehicles of cognition.

We do not Automatically Endorse our Internal Outputs

How trustworthy do we find our biological memory? Do we automatically endorse it or the outputs from our other cognitive capacities? I am willing to bet that Clark and Chalmers have at some point in their lives had reason to doubt the output of some part of their internal cognitive system. A hunch ignored, a visual illusion disbelieved, a recollection considered suspect, a sense of spatial disorientation supressed, a

confabulation discovered. Aware of conflicting accounts of past events, or informed about the shortcomings of biological memory and its constructive nature, do we always endorse all our recollections? I do not. The bar then is not so high for the endorsability of externally located content, if we are comparing it to internal content.

I do not wish to imply I do not trust any of my cognitive faculties, it would be rather hard to be a functioning human being if that were the case! Many of the outputs of the workings of our brains are automatically endorsed, in the sub-personal sense. I cannot but have the experience of *greeness* when I look at leaves, I cannot but automatically get the answer '4' whenever I think of '2+2', and when I try to recall what I had for breakfast a vision of fresh baked bread leaps to mind. But as informed and reflective entities that encounter counterfactual evidence that count against our senses, all our sub-personal automatic endorsements are ultimately open to assessment and judgement. I may never be able to doubt that I am having an experience of *greenness* if it seems to me like I am having one, but I may worry whether I am hallucinating the *greeness*. And maybe I'll accept my partner's recollection of what we had for breakfast over my own. We do not immediately endorse all information furnished to us by all parts of our natural cognitive system so there should be no reason to suppose that information from an external source should "more or less" be automatically endorsed to qualify as part of our cognition.

4.1.3 Easily Accessible

The Accessible criterion is "that information contained in the resource should be easily accessible as and when required" (Clark 2010, p.46). This suggests the system or

process has to work sufficiently efficiently, effectively, and speedily. This does not need be the case. Consider how we sometimes fail to remember something. Sometimes we walk into the kitchen and wonder why we have done so. Sometimes we cannot quite work out why a particular character in a film seems so familiar, and which films we must have seen them in before (cue the cognitive prosthesis of a quick search on the internet). We forget what we were just about to say and earnestly tell our interlocuter we had an interesting point, and the loss nags away at us until we finally recall it a bit too late, and discover it was something not quite so interesting after all. Sometimes we just cannot come out with the right word. For something to be part of a process or system, it does not need to be "easily accessible as and when required". That is a fine system design goal or property to have but is not characteristic of a system or extended process. The resource needs to have been accessed to contribute to a cognitive process, but whether it is easily accessible or not is a matter for when and how it is used, not whether it contributes to a process when it is used.

One could object by saying that these cases are unusual, that by and large our brain-bound cognitive resources are easily accessible, and so we might wish to reject those cases where it is usual for the candidate resource to be inaccessible. In response to this, I think the argument made via the neuron replacement teleporter device (see section 4.1.1) holds against this objection, though the critic can point to that being a case of limited *token* access but easy *type* access. Even so, if there were some part of (say) Otto's brain which could only rarely be accessed, we might want to make claims as to whether or not it was part of Otto's general cognition or part of his mind, in terms of membership of a system, but it would remain the case that when it is accessible, it

would be participating in and constitutive of the cognitive process, and the same can be said of external resources as well.

So, while it might be the case that generally our natural cognitive capacities are typically accessible, there are cases where they are not (e.g. when we cannot recall a fact), but while some of our capacities or resources might not be accessible at any given time, it does not rule them out from being genuinely cognitive when they participate in a cognitive process (Sprevrak 2009, p.13). There is no standard way of functioning that requires some resource that is used as part of a process to be accessible (or available) more generally. That most of our brain-bound cognitive resources are generally accessible (and available) is a contingent fact about our unextended cognition, not an essential feature of cognition.

In this Clark and Chalmers (1998) agree with me. They point out that the biological brain is at risk of its capacities becoming temporarily inaccessible, due to causes such as sleep, intoxication, and emotion. In these cases, when something is not accessible to wider cognition, then it is not participating in that cognition. But again, as with my wider argument, when a cognitive capacity is accessed it is participating in the process just as much as any other capacity, and therefore there is no grounds to say it is not part of cognition at that point in time.

4.2 Embracing Cognitive Bloat, the Extended Mind Extended

In what follows I will seek to argue the case for embracing cognitive bloat. I will take as my starting point that cases like Otto and his notebook *are* cases of extended cognition;

Otto has some set of his beliefs by virtue of being coupled with his notebook. Perhaps you already had this view[50], or perhaps I have persuaded you through the course of this paper.

So far, I have given a statement of the functionalist case for cognitive extension, defended it from some of its strongest critics, and considered some of the *reductios* levelled at the Extended Mind Thesis. I have also sought to persuade you that Clark and Chalmers' Availability and Portability Criteria are not successful in limiting the scope of cognitive (and mind) extension, and therefore a more liberal or 'bloaty' view of extension should be considered. Next, I will introduce the idea of 'Extended Digestion', to argue by analogy for the general principle of process-extension with a less contentious example.

[50] In a 2020 survey of philosophers carried out by PhilSurvey (https://survey2020.philpeople.org/) 51% of respondents said they accept or lean towards accepting the Extended Mind Thesis.

5 The Extended Digestion Thesis

"It is only for convenience (and from habit) that we think of organism and environment as separate; in fact they are best thought of as compromising one system".

—Anthony Chemro, *Dynamical Explanation and Mental Representations*

According to the Parity Principle, if some part of the world functions as a substitute for part of a cognitive process, then it is part of the cognitive process (Clark and Chalmers 1998). The Parity Principle was put forward about cognition, but there is nothing specific or special about cognition that confines the argumentative move to cognition. The Parity Principle can be applied to other processes to create an argument for process extensions. In this section, I consider the potential case of extended digestion. This discussion is intended to broaden and explore the idea of process extension, to cast it under different lights, to see what we make of it and what it means for the extended mind, and in general what it means for something to be part of a coupled system. Establishing a more general case of process extension strengthens the case for cognitive extension. Further, if process extension, including extension of bodily processes, is a general phenomenon, then the critic of the extended mind theory must make a case for cognitive exceptionalism when it comes to process extension.

5.1 Psychological Barriers

It is perhaps natural to find the Extended Mind Thesis and process extension more generally to be counterintuitive. It may even be the default view to think of processes as contained within physical boundaries. Dennett (1989) suggests that we are psychologically predisposed to think of the edges of our bodies as a boundary, but that this boundary is also flexible. He thinks that distinguishing self from other is a 'deep'

biological principle, founded on self-preservation. Having a boundary around oneself, allows one to fend off those that are other, and for organisms to preferentially allocate resources to itself. For example, the immune system distinguishes its host from intruders. We have a sense of our body, and of the bodies of others, and an aversion to not-us entering our boundaries. This is amply demonstrated by our readiness to swallow our spit or lick our blood from a pricked finger, but our general disgust and reluctance to drink a glass of water we just spat in, or to consume our own blood from a bandage or some other surface (Rozin 1987).

The borders of the perceived self are shiftable. This is starkly evident in cases such as somatoparaphrenia, where sufferers deny that one of their arms (usually the left) belongs to them (Vallar and Ronchi 2009). This can manifest in various ways, including referring to the arm as something dead or monstrous, and even asking for it to be removed (Vallar and Ronchi 2009). This shows that the sense of self, of ownership, of our bodies can contract. The sense of body can expand too. People can come to think of a rubber hand, virtual reality representations, and mechanical grabbers as part of their bodies (Carindali *et al*. 2021).

In terms of cognition, as with other processes, we are predisposed to consider its boundaries to be delineated by the borders of bodies, because that's part of how we view the world. Thus, there is a certain intuitive weight telling against the acceptance of extended cognition. This predisposition has advantages, such as being able to identify clear boundaries and persistence of entities. However, the world is filled with things that we cannot see and do not include in this visual joint carving. For example,

networks joined together by wireless transmission. Or consider the symbiote bacteria colonies that live within us, without which we would die. These are parts that belong to a larger whole that we do not readily recognise as belonging, but in certain ways they do. In much the same way we cannot see cognitive processes (only their outcomes exhibited as behaviour), or see the wireless transmission of information in a network, the activity of bacteria in our gut, or the invisible things that make a group or an organisation, they still participate in activities and processes that are not tied to one physical entity.

What this points to, is while we do have a physical border of 'skin and bone' to our bodies, which sharply delineates a certain distinction between us and not us, things are not quite so clear cut as that. There are ways in which what counts as us does not have to include everything inside our bodies (the bacteria for example), and there are ways in which things outside of our bodies can count as us. This is the case with our extended minds, where our minds extend out into the world. But this is not just some exotic philosophical claim that you would have to be crazy to believe, it reflects how various processes outrun an object's physical boundaries and is just one example amongst many. Next, I introduce the idea of Extended Digestion as a parallel case to extended minds, to help make the case that such cases of extension are natural and everyday, and not something to be denied.

5.2 Digestion

Where is digestion? A typical answer might be that digestion is a process carried out by the body. It is a process that takes food and renders it into a form that our body can

absorb and make use of. Typically, digestion is thought of as taking place in the stomach, using acids and enzymes, and in the intestines. Digestion also takes place in the mouth, via the mechanical breaking up of food from chewing, and the breakdown of starch by an enzyme in our saliva called amylase (Meadows 2008, p.11 and Travers 2019). Mechanical digestion occurs elsewhere in the body too, and not just as a way of transporting food matter between steps of the digestive process. For example, muscular contractions in the stomach churn food with gastric juices to produce chyme, which is then passed into the small intestine for further digestion.

I will argue that digestion can be extended. This means that external things in the world can form part of or enact part of our digestion. Sometimes, some of our digestion happens *out there*, in the world. Digestion is not limited by the borders of skin and bone. And as with digestion so with cognition.

First, it is worth considering some of the varied and exotic ways other animals render potential food into a form that their bodies can absorb. Some species of Starfish extrude their stomachs from their bodies to begin the process of transforming their prey into simpler elements before drawing in the resulting slurry (Anderson 1954). Reduviids, a family of insects that prey on other insects, inject their prey with a saliva that begins to break down their tissues in an act of 'extra oral' digestion (Kumar and Sahayaraj 2012), before ingesting it. These cases involve parts of the digestive process happening outside of the animals' bodies.

Digestive action, I suggest[51], demands spread of *digestive credit*. If, as we confront some task, a part of the world functions as a process which, *were it done in the body*, we would have no hesitation in recognising as part of the digestive process, then that part of the world *is* (so I claim) part of the digestive process. Digestive processes ain't (all) in the body!

At this point a critic could object that while these may be examples of chemical digestion-like activities carried out by other animals, it is not the same as human digestion, which is restricted to internal digestive processes. I will seek to sway these objectors and convince them that human digestion is not confined to the body.

5.3 From Plate to Mouth

Let us imagine the case of Chompers. Because he lost his original teeth, Chompers now uses a set of false teeth in their place. The false teeth look the same and work the same as his original teeth. When Chompers wants to consume food, he cuts it up into smaller pieces through a normal chewing action using his false teeth. With his false teeth he can eat all the things he used to be able to eat with his original set of teeth. We can say that Chompers behavioural competence, when it comes to chewing, would be degraded if we took his false teeth away from him. Despite using an external artefact, Chompers is recognisably carrying out a digestive action by chewing his food with his false teeth. While we can point to differences between regular chewing and Chompers chewing with his false teeth, we can recognise at a coarse level that it is the same causal role being carried out.

[51] Following Clark and Chalmers (1998).

Sometimes, Chompers takes his false teeth out of his mouth and uses them to *chew his food while it is still on his plate*. When he does this, he is cutting the food up into smaller pieces, an act of mechanical digestion, outside of his mouth, outside of his body. If Chompers' teeth were in his mouth, we would have no hesitation in recognising them as part of his digestive process. Outside of his mouth they are performing the same functional role in the digestive process, and therefore (I claim) they should be considered as part of his digestive system carrying out the digestive process.

To be sure, when Chompers holds his teeth in his hand and uses them to chew up the food on his plate, it is not quite the same as if he had been eating normally. There is no saliva to start breaking down starch for one thing, but details such as this are not central to the case. Chompers is carrying out the mechanical process of chewing with his false teeth, even if additional elements are absent, and in any case, Chompers could do something like sprinkle amylase over his meal. The essential characteristic of rendering down food via chewing is still taking place, even though it seems peculiar to call this activity chewing or digestion, because it is not happening inside his mouth.

Chompers' false teeth are quite close to the case of having natural teeth, and so their use is not as radical a case of extended digestion as Otto's notebook is for an extended mind, but we can make the case of Chompers a more radical case of process extension. Imagine Chompers loses his false teeth and must rely solely on a knife and fork to cut his food up. The same kind of mechanical digestion process is taking place as when he uses his false teeth. Food is being turned into smaller chunks of food that are easier to

swallow and are better prepared for chemical digestion in the stomach; we have the same input to output mapping for the process.

Chompers and his cutlery can be said to be two separate components that constitute a new extended digestive system. If the cutlery is taken away from Chompers his behavioural competence at eating and digesting will decrease. Missing his teeth and deprived of his cutlery he might even starve. The contribution of the cutlery has a 'direct' impact on Chompers' behaviour, they play a 'crucial' role in the 'here and now' for digestion.

We can run the same kinds of moves with Chompers and his cutlery (or false teeth) for extended digestion, as Clark and Chalmers (1998) do with Otto and his notebook for extended cognition. We can say that because Chompers is so reliant on his cutlery, he always keeps them with him, so that they are reliably available and accessible whenever they are needed.

The case of Chompers is instructive, for while we do not all have replacement teeth, many people around the world use external artefacts to carry out the initial stages of digestion prior to putting food in their mouth. Our digestion is not limited to the inside of our bodies. It can happen out there in the world.

5.4 Digestive Bloat

Extended Digestion faces similar boundary challenges as the Extended Mind. We may be in danger of overextending digestion, but we also face the danger of being too conservative in how we think about processes. We should want to avoid thinking that

processes are always contained within physical boundaries and explore how far they extend.

Are we left with digestive bloat? If cutting food into smaller parts—whether via our teeth or our cutlery—is digestion, we must consider other cases of cutting. Is the chef in the restaurant kitchen carrying out part of our digestion when she chops vegetables as part of preparing our meal? Is the factory that turns fruit into a smoothie part of our digestion?

I think the answers to these questions is a qualified yes. These are cases of extended digestion. However, we need to draw out the ambiguity of what we mean by 'our' or 'my' digestion. When Chompers takes out his teeth to chew the food on his plate it is *his* digestion. The organism that is Chompers is acting to mechanically break down the food, which as part of a longer process will lead to the nutrients being absorbed by his body. In this case, it is Chompers interacting with the world, and it is his interactions that deserve the digestive credit.

When someone else cuts up Chompers' food for him, it is still digestion, and it is even part of an extended process of digestion *for* Chompers. However, it is not Chompers doing the digesting himself. It is digestion for him, but not by him. Chompers is not interacting with the food, until it comes into his possession, so whatever is happening to the food it is not part of Chompers' digestive process. Chompers is not coupled to anything involved in the food when it is being cut up for him. One could say that Chompers *is* loosely coupled to the person cutting up his food. There could be a loopy two-way communication between that jointly governs the digestive activity but

granting this would establish that there was an extended digestive process spanning the two people, but not that the whole process is Chompers' process.

In a case like this, the other person is changing the input to Chompers' digestive system, is it is not an extension of his digestive system such that there is a new functional arrangement to carry out the process. In the same way, this is why Otto and his notebook is a candidate for cognitive extension, but Otto reading a random road sign is not. It is the difference between engaging with the world in a way that offloads some portion of a process across external objects, and mere input.

The claim I would like to advance is that for something to count as Chompers' digestion, it needs to be in some sense carried out by him (as an organism) or under his direct control. It needs to count as a (partial) substitute for part of one of his bodily processes, not merely being an input or something effecting change in the world. The chef preparing food may well be involved in an extended chain of digestion, and she may well be partially digesting Chompers' food for him, but it is not Chompers doing it. This remains the case even if the chef is directly asked or ordered by Chompers to cut his food.

What I am trying to draw out is that we should recognise continuous processes extending out from our bodies and our control. We view the world in terms of objects doing things or containing processes (what Ross and Ladyman (2010) refer to as the container metaphor), but processes also flow across objects, such as in the case of information, heat, or kinetic energy transfer. We may have processes contained inside of us, but there are also processes flowing in and out of us. Establishing ownership over

part of the process (*my* digestion or *my* cognition) does not split up or end the process, rather it establishes ownership and influence over part of the overall process.

With this view in mind, we can tackle the problem of digestive bloat. Is premastication of food for an infant digestion? Yes, but it is owned by the one doing the chewing. It is not the infant causing the digestion to happen. If, as in the case of Chompers, using cutlery to cut food is digestion, is using a blender? I do hesitate to say yes, because if pushing a button to blend is digestion, why isn't pushing a button to instruct a swarm of robots to construct a factory that will start blending food on your behalf count as digestion? I think the answer here is that in the case of the blender, the operator is typically in direct control, making the blending happen. In the case of someone using a swarm of robots, then there is delegated or derived control, so it does not count as their digestion. The two-way loopy interaction between the owner of the robot swarm relates to carrying out a process that is about instructing the robots what to do, however it is too indirect and not coupled to the blending that eventually takes place in the factory.

The question of ownership and control is a worthwhile topic to pursue, but I have only touched upon it lightly here. I wish only to establish that there is external extended digestion (and draw parallels to the case of cognition), and that it is common and widespread. Who it belongs to, once it is outside of the borders of skin and bone is a broader question and in part dependent on which particular interests we have in setting out ownership schemes.

5.5 Summary

I have argued that digestion can be extended into the external environment, and that items like knives and forks can be used as a substitute for part of the mechanical action of internal digestion. The case of Extended Digestion shows us that the underlying argument for the Extended Mind can be applied more generally to other processes. In turn, this shows that mind is not a special case, and a whole range of processes enacted by our bodies may extend into the external world. As such we should be more open to accepting the Extended Mind thesis and its implications. It also implies our actions may constitute parts of external processes of which we are not even aware. This all serves to highlight the nature of processes and how they need not be constrained by the boundaries of physical objects. Even though the claim that things external to us are carrying out parts of processes normally carried out organically by our bodies is a strange one, any distinction between internal and external, or real versus artificial, are distinctions about human-centric interests, not about the physical functions or processes taking place.

6 The Extended Mind, Revisited

6.1 Limiting the Mind

I have argued in defence of the Extended Mind Theory. I have also argued against the standard version of the Extended Mind Theory, which places significant constraints on how the mind can extend into the world. I have argued in favour of cognitive bloat, seeing it as a reflection of how things are rather than an absurd and undesirable outcome to be avoided. In support of this I have tried to make a case via the example of Extended Digestion that instances of the Extended Mind are just as mundane as other extensions of processes, and therefore we should be more accepting of cognitive extension.

Some may see my efforts as ultimately destructive rather than productive. I have strengthened the case *for* cognitive bloat, and there a risk of smearing our idea of cognition out in the world so thinly as to make it trivial. To make amends, somewhat, I have a modest proposal on how we should think about cognition a little differently, which I hope will be productive.

This view is this. Cognition is a natural process. The most salient and interesting cases of cognition that we are aware of take place in brains. But cognitive processes are not limited to brains and can extend out there into the world. Again, this is a normal occurrence, just as with other processes like digestion. Cognition, or bits of cognition, happen 'out there' far more commonly than we might suppose. We need to recognise this. The main questions of debate that arise from the Extended Mind Thesis are "is this cognitive?" or "is this part of X's cognition?". I think these are not the most

productive questions to be asking, and as I have been arguing I suspect in most instances of debate the answer is simply; yes. What is going on is cognitive because it is doing cognitive things. Otto really is retrieving beliefs. The Tetris player really is mentally rotating shapes. Part of the causal role is fulfilled, and the activity occurs. The more useful question which I think should be asked is “whose cognition is it?”.

This question is important because the view of extended cognition I am proposing is a wide view. It actively endorses cognitive bloat. The picture I want to create is that within our bodies, and particularly in our brains, there is a great deal of cognition going on. We, and creatures like us, are a concentration of cognition in the landscape. But cognition is not just within us. When we are coupled to an appropriate part or activity, it leads to cognitive processes extending across the parts. I imagine if we could directly see cognition, we would be glowing centres of light, but there would be strands of light flowing between us and across the landscape. Some cognition stays within us. Some starts with us and goes out into the world, some come from the world into us.

If cognition is all over the place like this, the more pressing question becomes “whose cognition is it?”. I propose the following criteria as a replacement for Clark and Chalmers’ Availability and Portability Criteria: if something is participating as part of our cognitive process, and we have control over it, then it is part of our cognition. Or, as Dennett (1989) puts it “you are what you control and care for”. I will focus on control in what follows.

6.1.1 Control

Here, I set out what I consider to be a useful way of thinking about cognitive extension. In this, I am accepting radical cognitive extension, unrestrained by the Availability and Portability Criteria and concerns about Cognitive Bloat. Cognitive extension occurs when our cognitive processes are coupled to an external entity via our loopy sensorimotor interaction in such a way that both the internal and external parts jointly govern the relevant behaviours. The external entity in some sense participates in the cognitive process as part of a combined causal system.

This view leads to not only accepting cognitive extension, but that cognitive processes which we are part of or participate in can extend widely, such as in the case with Otto and Notto the conjoined twins, using a library, or in the case of Chompers and digestion with the blending of food by a chef to form part of a digestive processing chain. Rather than placing artificial human-centric limits on what might count as cognition or cognitive extension, my focus is on placing human-centric boundaries on what should count as *our* cognition. The idea is to identify those external cognitive resources that we are incorporating into our cognitive processes by exerting control over them. Along with control come notions of ownership and responsibility, which I will also discuss briefly.

The answer to “whose cognition is it?” when it comes to cases of cognitive extension is that it belongs to whoever meets the following criteria:

1. They have meaningful control over it: the extension is initiated or accepted by them and is not subject to undue interference by others; and
2. The control is goal-directed: the extended process is used in alignment with their goals and agency.

To control something, we need to exert meaningful influence over it without that influence being strongly contested. Control need not be absolute, and what counts as meaningful control may vary across contexts. My heart is mine. I can control it indirectly, for example by doing some star jumps to make it go faster, but I (as a person) do not have direct control over it, though the wider organism of which 'I' am a part does control the heart.

To control an external part of a cognitive process, as well as needing to be appropriately coupled with it, the extension needs to have occurred due to the agent's intent or acceptance of using it to achieve some cognitive goal. The way in which it participates in their cognition must be under their influence, such that the agent is driving the external entity's contribution to the process. The relationship needs counterfactual dependence, such that there can be different outcomes if the agent acts differently.

Control can be mixed or contested. For example, there could be competing legitimate claims to the same cognitive resource, perhaps two people jointly own the same notebook which they use as part of their extended cognition. There will be times when an agent is in control of an external resource, and times when they are not. Otto may lend his notebook to his friend Clara for instance. He no longer has direct control over

the notebook, even though it is still his in the sense of property ownership. In this instance Otto retains a form or social or legal control over the cognitive resource, but he does not have direct control over it.

There are different types of levels of ownership. As a rough pass, I propose that a First Order 'mine' are the parts that I'm doing with my body (thinking about stuff), the Second Order are those parts that I am directly making happen out there in the world (writing in a notebook), and the Third Order 'my' cognition, which is much less mine, are those parts being done at my direction or with my property. There is cognition, it extends and couples all over the place, and I have varying degrees of 'ownership' and 'control' over it.

These different orders of ownership indicate how much control we have over a thing. Admittedly, in the case of our own bodies we have a form of mixed control, in that many bodily processes happen without our involvement, or even against our wishes, but we recognise our bodies as our own. The matter at stake here is about which *extended* cognitive processes count as ours.

There are also cognitive resources available as public goods or collective commons, or simply available and unowned by others. Examples include information resources on the internet or a library. When I read an online encyclopaedia article, it may contribute to my cognition, but it is not *my* cognition, because I do not have control over it. At any moment the owners of the article could delete it or change its content. But I can acquire it, by making a local copy, and therefore make it mine, or I may be able to make changes to it and so exert a degree of control over it, making it a Third Order resource.

Here I am not merely asserting that control matters for whether something counts as our cognition, rather I am pointing to the parallels as to how we already decide who a body or a notebook belongs to. Cognition, being an activity rather than an object is different to bodies and notebooks, but a similar view can be applied to other activities, such as who a gesture, a song, or a dance belongs to. What matters here is that the agent has control over whether and how an external entity participates as part of their cognitive processes. It is this control and active incorporation which makes it part of *their* cognition.

7 Doorstops

"Sometimes I'll get one of the blokes to sit down against a door so that nobody can get through. At that point, he might as well be a doorstop"

- a soldier, speaking to the author

In this section, I aim to make the case that i) functional substitution can be coarse grained, so long as the essential causal role of a function is enacted, ii) the functional substitution does not need to be stable or long lasting for it to be successful or to support process extension, and iii) there is much greater scope and range of candidates for cognitive extension than we might suppose, because we are somewhat misled by our human-centric perspective.

Part of what I am doing in this section is highlighting the fundamental flaw with the Availability and Portability Criteria. The criteria may well be doing useful work, but the work they are doing is not to distinguish whether something is a cognitive extension or not. The basic underlying point that I wish to make is that whether these criteria are met or not, it makes no difference to whether an external entity has functioned as a substitute for part of a cognitive process. The same cognitive activity occurs, whether the criteria are met or not. The Availability and Portability Criteria are intended to pick out the *more modest and psychologically plausible seeming cases* of cognitive extension (Clark 2010, p.46). Beyond this, no justification for them or why they should be considered is provided (Clark 2010, p.56 and Sprevak 2009, p.12). They are a rhetorical move to encourage acceptance of the extended mind thesis. The criteria are not about identifying whether or not the substitution has occurred, they are about identifying extension cases that seem relatively close to our regular cognition. That

may well matter for some questions, but not as to whether cognitive (or mind) extension has occurred or not.

I advance my case with regards to broader acceptance of functional substitution via an analogy involving doorstops. While doorstops enacting a doorstopping function are markedly different to objects being involved in an extended cognitive process, the analogy serves as a simple model to make the general case.

There are many objects that we recognise as doorstops. But there is a larger set of objects that could be used *as* doorstops, that may be just as suitable or more suitable as doorstops than 'official doorstops'. We may generally consider official doorstops to be such based on a range of factors, including convention, recognisable features, and the intent behind their manufacture. However, these criteria are not as stable as we may ordinarily suppose. Official doorstops can stop being doorstops, and non-doorstops can become doorstops when thought of in the right way.

While we may have good reason to draw a sharp distinction between official doorstops and other objects that are suitable to be but not considered to be a doorstop, we should not neglect that both sets of objects have important and stable properties in common. When confronted with two sets of objects equally suitable to be to be used in the doorstopping role, but only one set is considered to be doorstops, what makes them doorstops is what we bring to the picture, our human perspective and interests. That perspective and those interests can change, and we may overlook important facts about objects in the world, including their potential causal roles should they be coupled to other objects in an appropriate way.

In what follows I seek to weaken the distinction between what is ordinarily considered to be a doorstop and objects which are not considered to be doorstops but could be used as one. What counts as a doorstop can change across contexts, and no particular objects have the priority as to which could be official doorstops. None of them have a privileged position of being 'the real thing', because they are just as apt in themselves to qualify as 'the real thing'. There are merely different contexts in which objects with the currently appropriate criteria can be considered doorstops. That is not to say that just any object can be a doorstop, as there are still criteria that need to be met in order to be a doorstop, but (perhaps for entirely pragmatic reasons) our day-to-day ideas about what objects are or can be doorstops is too strong. I think we should be open to thinking more broadly about what objects are or can be doorstops.

The parallel I am seeking to draw is with cognitive systems. As with doorstops, so with cognitive systems; in picking out what things are normally recognised as fitting within a particular functional role, we must not neglect the less stable, unusual configurations, or good-enough substitutions that can realise the relevant causal role. Just as there are potential doorstops all about us in the world, there are also potential parts of cognitive systems.

7.1 The Humble Doorstop

Consider the humble doorstop. A doorstop is an object that stops a door from moving, normally but not always to keep it in an open position and thereby prevent it from closing. Typically, this stopping is for the purpose of keeping the doorway free of

obstruction, to allow frequent movement through the doorway without having to open the door or hold it open.

In use, a doorstop is placed by a door to be stopped, and when placed correctly it stops the door from moving. There are two main variants of doorstops. The first type works by being wedged between the door and the floor, so that that door cannot swing freely, and so is stopped. The second is a mass placed as a barrier between the door and its closed position, to prevent the door moving further. This type of doorstop work because the force acting on the door to make it swing shut is not great enough to shift the mass of the doorstop.

There are a wide range of doorstops. The wedge type are limited in form because it is their shape that allows them to stop a door. The get-in-the-way type of doorstop come in a wider variety of forms, for example they might be a small weighty (imitation) stuffed animal such as a small dog. All that is needed from these doorstops is to have sufficient mass, and sufficient friction with the floor (a wheeled doorstop might just roll out of the way!) to ensure they do not move when the door pushes against them.

Other schemes and arrangements for stopping doors exist. They all stop a door from moving by itself. In what follows, I will focus on the objects that we actually refer to as doorstops; the wedges and the placeable masses. More specifically, the types of doorstops that I am talking about are typically simple objects, with no moving parts, can be readily moved by an adult, and do not require any modification to the door or the door's surroundings to be used effectively as a doorstop. They all work by putting them on the floor up against the door.

7.2 Carving Nature by Human Interest

The world is filled with things that we try to make sense of. Many areas of enquiry are concerned with carving up nature into schemes that make sense. Sometimes we'll carve up nature in keeping with what we discover about how things work, even if they conflict with our human-centric view of the world, such as with the elements and the periodic table, or the shift to a heliocentric model of our planet's relationship with the sun. Often, as is the case with doorstops, we carve up nature based on our human-centric view of the world, based around our interests rather than what they are.

Why do we agree that the objects that we call doorstops are doorstops? We do so chiefly due to experience of these objects, with our practical use and interaction with them, and from what others tell us. These are human reasons and human interests. We see objects advertised for sale as doorstops, people tell us they are doorstops, and we see them in use as doorstops. When we see doorstop-sized objects for sale in the Doorstop Shop, we believe that they are doorstops, else why would they be for sale in the Doorstop Shop and labelled as a doorstop? For these sorts of reasons, we consider these objects to be doorstops.

From this everyday experience of doorstops we have an appreciation of their features that enable them to stop doors, and also of their canonical appearance. These features form part of our concept of a doorstop, which we use to identify other objects as doorstops, or to decide that some other object is not a doorstop. If I handed you a small piece of cheese, and said that it is a doorstop, you would likely not agree with me, because the cheese does not share the typical features of a doorstop. Doorstops

need to be capable of stopping a door. They tend not to do things like grow mould on them or get eaten by people. Admittedly, cheese does often come in a wedge shape, but given its low rigidity, tendency to smush when subjected to force, and to emit an odour, cheese generally makes for a poor doorstop. Trying to use it as such would also be a poor use of good cheese.

There are a set of features that make an object physically suitable to stop a door, and so potentially qualify to be a doorstop. Distinct from these, there are other relevant properties. There are objects *meant* to be doorstops. They have been designed to be a doorstop and have been made to be a doorstop. As with the physical features, these properties can change but can also prove to be stable. Consider an object that is intended to be a doorstop, but due to some flaw or accident it is no longer capable of stopping a door. It is a broken or defective doorstop. Yet, despite not being able to use it as a doorstop, we may still say it is a doorstop. In this case when we call it a doorstop, we are calling out its causal history, its intended purpose, and the similarities it shares with working doorstops. But we are not calling out its current purpose—the uses to which it can be put—as it can no longer realise the function of a doorstop. If I were to say, “Pass me the doorstop”, and only the defective doorstop is at hand, you might be right to pass it to me. But consider a different scenario in which there are two doorstops close by to a door, one known to us both to be defective, and the other to be an effective doorstop. If you needed to stop a door and said, “Pass me the doorstop”, it would be a mistake to pass you the defective doorstop. We recognise that in some scenarios, the defective status of that doorstop will exclude it from the set of objects we mean to pick out when we refer to doorstops.

What this indicates is that sometimes we carve up nature by human interests and viewpoints, and this is also the case with doorstops. But what an object could be used for also matters, over and above these human-centric considerations. And we can mean different things by function; an object's intended function, the functional roles it can realise, the role an object has by virtue of its relationship with other objects (e.g. being part of a broader system) whether it is working correctly or not, and the current function it is carrying out. It is to this kind of idea, of the causal role that something performs or *could* perform (Couch 2025) to which I appeal.

7.3 Functional substitution

A typical doorstop is capable of being used to fulfil the functional role of stopping a door moving. We can specify this functional description further by saying that doorstops fall within such and such a size range, and such and such a weight and mass, and so on, to be usefully used as a doorstop. Using this description, we can pick out which objects are doorstops.

There are some objects that fit the above functional description of a doorstop and are also intended to be doorstops. This set is much smaller than the set of objects that could be used as a doorstop but are not intended to be a doorstop. I have seen bricks and chairs used as doorstops. The claim I am arguing for is that these 'makeshift doorstops'—such as bricks and chairs—are as much a doorstop as the regular kind, once you take away what we bring to the picture, our human interests and perspective. While one could say if you take away human interests, then doorstops are no longer a category of thing, but this is the view I am seeking to challenge. The point I'm making is

that there are range of objects that could be used as a doorstop, and when they are used in the doorstopping role we should grant them the status of being a 'doorstop', even if that use is *ad hoc* or is an unusual for that particular object.

I am not seeking to overturn how we recognise and categorise objects, but I wish to make an important point about how we consider objects. That point is this; even if we are not considering the claim that a random brick *is* a doorstop, we should recognise that it *could be*, and the only significant difference between it and a regular doorstop is what we bring to the picture, of our deciding whether it is a doorstop or not and using it as a doorstop or not. Given we could decide at any moment to use a random brick as a doorstop and thereby change its status we have cause to consider just what all the objects around us are, and to what extent we are surrounded by these potential doorstops. Our assigning a functional role or identity to an object is perhaps based more on factors such as convention, tradition, and intuition than we might like, causing us to not fully recognise the object and the range of functional roles it can potentially realise.

While I think the position that every object that is functionally-suitable to be a doorstop could be a doorstop recognises a truth, it is only a partial truth. All these functionally-suitable objects are only that, *suitable* to be a doorstop. They are potential doorstops. Doorstops are human things. More than fulfilling the functional role, some human or other agent must think of the object as suitable to stop a door. This gives us two sets of properties; those that make the object physically suitable to stop a door, and additional relational properties that lead to us considering it to be a doorstop.

‘Official’ doorstops meet both sets of criteria. Defective doorstops only seem suitable. Objects that are physically suitable but are not considered suitable—whether through their appearance, convention, or simply not being thought of in that way—are potential doorstops.

It is this second set of properties, of how we consider an object, that I want to put under some scrutiny, and make the barriers between an official doorstop and potential doorstops more porous.

7.3.1 The Traditional Brick Doorstop

Consider a brick that has been used to stop a door in an old building for centuries. Everyone familiar with the building knows the brick is used as a doorstop. This longevity and history of use as a doorstop, is part of what establishes it as an official doorstop. The brick doorstop has become part of local tradition. It fulfills the causal role of a doorstop well; never has the door in the old building closed by itself when the brick has been properly employed as a doorstop. So famous is this traditional brick doorstop, that tour guides stop to discuss it, and the building’s gift shop sells replicas of it. Any visitor can buy a replica of the old brick doorstop from the gift shop, to use as their very own doorstop.

Let us further suppose that the brick doorstops in the gift shop are made in a small factory that only makes bricks for this purpose. Every brick made by the factory is meant to be a doorstop and is shipped to the gift shop to be sold as a replica of the tradition-soaked old brick doorstop. This means that the bricks made to be replicas of

the brick doorstop are doorstops both due to human intentions and considerations, and for their suitability to realise the causal role of a doorstop.

A Confusion of Bricks and Doorstops

When delivery trucks full of what people would ordinarily call bricks drive down the road in this town, we have an interesting case of uncertainty. Ordinarily when seeing such a cargo we would assume that they are bricks, made to build things, and on their way to some depot or construction site. In this town, it is just not possible to tell whether they are bricks, or rather whether they are intended to be used as construction-bricks. The truck could turn left to the construction site, in which case the future use of the objects will be to make houses. Or the truck could turn right, to the old building and its gift shop, in which case their future use is as doorstops. When we know about the gift shop and its replica doorstops, and do not have further knowledge about who ordered the bricks, which manufacturer they came from, and so forth, we are not able to say whether the objects in the truck are bricks or doorstops. While there is a fact of the matter, in terms of their intended purpose and similar considerations, in terms of functional use it does not matter which purpose they are put to.

The picture is further complicated in the case of the sleepy truck driver. Sometimes he has a job to deliver bricks to the construction site, sometimes he has a job to deliver replica doorstop bricks to the gift shop. Some days the sleepy truck driver forgets which job he has and being unable to tell from just looking at his cargo he flips a coin to decide who he will deliver to. This has been going on for months. Now we have a

situation in which two sets of objects, made for different purposes, will be used for one purpose or the other based on pure chance, a literal flip of a coin. The intent behind their design and manufacture has no bearing on whether they end up being thought of as bricks or doorstops, whether they are used to build houses or stop doors from moving. It all comes down to the coin toss whether the objects will be bricks or doorstops.

Unfortunately for those who wish to track which objects are really doorstops and which are not, a mischievous local often sneakily swaps the bricks in the gift shop for standard bricks from a local construction site. The result is that some construction site bricks get sold as replica doorstops to tourists, and end up being used as doorstops, and some bricks meant to be replica doorstops become parts of walls. The result is that despite the intended purposes and other human-centric considerations relating to doorstop-bricks and construction-bricks, there is no barrier to their functional substitution. Whether a particular brick was meant to be a doorstop or not does not impact whether it can be a doorstop.

The case of the shop switcheroo has become even more interesting since the sleepy truck driver started delivering to town and tossing a coin to decide where he would deliver his cargo. We now have objects that were made to be used to build houses, intended to go to a construction site for that purpose, but delivered to a gift shop based on a coin coming up tails. These objects were mistaken for replicas of the famous old doorstop, packaged up as such, and displayed proudly in the shop window. For weeks, customers and passers-by admired these objects and thought about using

them to stop their own doors. Before one of them was sold however, the mischievous local swapped it for a brick from the nearby construction site. Unknown to this prankster, what she was using as a fake replica doorstop was originally built in the replica doorstop factory to be a replica doorstop. She has in effect restored some of the misdelivered objects to their intended destination. We have a confusion of bricks and doorstops, their intended purpose unknown, and perhaps in some cases no longer knowable.

Brick or Doorstop?

The question arises; which of these objects are bricks and which are doorstops? On one view, people are stopping doors with bricks and building houses out of doorstops. We can track the journey of the bricks and the doorstops as they are made, delivered or misdelivered, swapped as a prank or not swapped, and say things like “This is a brick not a doorstop, even though it was sold as a doorstop in the gift shop, and now holds open a door”.

Another view is that the objects have a set of physical properties that make them suitable to realise a range of functional roles, and that there is no further truth to what they are outside of our interests and perspective on the matter. Yet, our interests can change even while the objects stay physically the same. In the story of the traditional brick doorstop the objects are equally meritorious of being a construction brick or a doorstop, and which they ultimately become depends on the use to which they are put. Our interests are important, and sometimes the causal history of an object will matter, but generally they will have no bearing on what uses the object could be put to.

This is the view for which I am arguing. A given object may fulfil a multitude of functional roles, including being a doorstop.

This means that every loose brick in the world is a doorstop-in-waiting. This might not be enough to say it *is* a doorstop. Conversely in some contexts it may not be enough that something has been designed and made to be a doorstop to say it is a doorstop. We have two sets of criteria for what makes something a doorstop, the first are the properties of the object that make it suitable to enact the causal role of a doorstop, the second are the other properties about the object relating to matters such as our perspective and its causal history. How we treat something might sometimes be driven by the first set of properties, and other times by the second set of properties. However, it is the first set, the degree to which an object can fulfil a causal role which is what is relevant as to whether it can be used in that role.

What this talk of a confusion over replica doorstop and construction bricks points to is that the functional identity of objects is less stable than we commonly suppose. Some objects, such as bricks, could become a doorstop in a moment, providing they are appropriately coupled to a door. This is the case at least from an objective perspective without what we bring to the picture with our intentions and beliefs. The same view should be applied to cases of cognitive extension or (coarsely considered) extended functional role substitution. Our perspectives matter, but just because something does not seem like a ‘true’ constitutive participant in a cognitive process should not get in the way of recognising its causal role in the process.

7.3.2 No Difference That Matters

Our perspective is important. But we must not overlook that it is not the full picture, and it may hide or gloss over important features of the natural world, and (in the case of doorstops) overlook, underplay, or be blind to the range of functional roles that something could fulfil. If a particular brick can be a doorstop, then every other singleton brick can be too. The only thing stopping them from being an official doorstop is human perspective and intent. That matters, but it does not change the reality of what things have the potential for.

The consequence of all this is that the world is filled with objects that are suitable for stopping a door, in the same or much the same way as official doorstops, whether we consider them to be a doorstop or not. They might not count as doorstops in the same way as official doorstops, but in terms of their potential to fulfil the functional role of a doorstop they are a kind-of doorstop. They are potential doorstops, meeting all the necessary causal role criteria for being a doorstop, making them of a functional kind with doorstops, even if they do not (currently) meet the criteria for being of a social kind with doorstops.

This distinction is important, because these functional doorstops could realise the doorstop function just as well as an official doorstop. When they enact the function, they stop a door just as a regular doorstop does. As a result of being placed in front of the door doorstopping takes place. We can add as many restrictions or caveats as we want as to whether they are official doorstops; whether they're reliable as doorstops, whether they are typically used as a doorstop, whether they're available and

accessible, whether they're endorsed as a doorstop, whether they're recognisable as a doorstop, whether they were designed to be a doorstop, and so on. But none of these things have a bearing on their inherent capacity to be used in the doorstopping role, and there is no material difference in the doorstopping that occurs when they are used in that way. Whatever the considerations, the door is stopped, whether by a doorstop, a brick, or a chair.

I hope that by this point, this line of argument seems straightforward and agreeable. Of course, a brick could be used as a doorstop, and yes, while we would not normally suppose a brick is an actual doorstop when it used as such the end result is indistinguishable (or close enough) from the regular kind of doorstopping. Putting aside cases such as a history of long use, or doorstops made to look like bricks, we might want to draw a line and say that a brick is not *really* a doorstop. But when we make this claim we are picking out things about the brick other than its potential to be used as a doorstop. We are saying something like "this object was not meant to be a doorstop, it was not originally brought here to be a doorstop, when people look at it, they will say it is a brick and ask me why I am using a brick as a doorstop". While this is likely true, it is a distinct matter from it being used as a doorstop and enacting the function of a doorstop suitably well. It is actually stopping the door. We are only warranted to say it is only stopping a door *like* a doorstop when we are taking a human-centric view, picking out a social kind. Even so, the brick would still be stopping the door *as* a doorstop functional kind, no different to an official doorstop.

In the same way, we should say the same sort of things when it comes to cognitive extension. Much could be said about whether Otto's use of his notebook makes it constitutive of his cognitive process, whether he has certain beliefs at least in part by virtue of the content of the notebook, whether conjoined-Notto can participate in conjoined-Otto's cognition, and whether your cognition could extend to a library. We should consider these cases in the same way we consider using a substitute item as a doorstop, in terms of fulfilling the relevant role when a brick is used to stop a door it *is* a doorstop. We might not call it a doorstop, we might have wider considerations that mean we should not consider it a doorstop, yet the door is stopped. The object in question is fulfilling the door-stopping function successfully, whether we consider it a doorstop or not. I am claiming the same is true in cases of cognitive extension. There might be a great deal of important things to say about whether it is really cognition or not, or whether the candidate extension seems plausible under conditions like the Availability and Portability Criteria, but that has no bearing on the object actually fulfilling part of the cognitive process. What matters is how the external object is integrated into a cognitive process, not what it is, what it was meant for, its history, or how plausible it seems.

It can be questioned to what extent we should accept a wide view of extension and whether we should include so much counter-intuitive and explanatorily redundant cognition in the world. Part of the pressure here is to wonder why we should be interested if there are so many things that can be potentially used for cognitive extension but are not. There are two responses I have to this. The first is simply that it is interesting and tells us (or reminds us) of their potential for use as part of an

extended cognitive system, even if there are no practical or epistemic consequences. The second is that as we could potentially extend our cognitive processes over them, that potential has relevance to future cognition. Consider computer storage devices not connected to a computer system. Most of the time they are redundant, but they have the potential to be connected and to be used, enhancing what the computer can do, and because of this we should be open to the idea that the various potential extension-artefacts could become explanatorily relevant.

8 Where is my Mind?

The Extended Mind Thesis leads to radical cognitive extension, to cognitive bloat, whereby our cognition can be extended widely to things such as libraries, the internet, and other minds. In response to this counterintuitive and absurd seeming outcome critics have raised objections about whether cognitive extension could occur in practice, whether the resulting activity should count as cognition, or have reached the conclusion that functionalism is likely false.

My response to the worry of radical cognitive extension is to argue that we should accept it as an interesting implication that tells us about how cognition and processes in general work. Cognitive bloat examples do not suggest that the Extended Mind Thesis is wrong, they show us that how our cognition works as we interact with the world is different to how we are predisposed to think about cognition.

In the course of this paper, I have set out to achieve several aims. I have set out the Extended Mind Thesis and defended it from some prominent objections. In doing so I have argued that coupling always entails constitution, and the point which matters is whether the newly constituted system can be considered cognitive or not. Taking up this challenge, in responding to Adams and Aizawa's push for a 'Mark of the Cognitive', I have argued that the answer to whether something should be considered cognitive or not should be considered coarsely. This has two main components. The first involved drawing attention to the need to take a relatively coarse approach to functional definitions of mental states, focusing on what is essential to the relevant mental states and their causal role in cognition-dependent behaviour. The second, via the analogy of

doorstops, sought to emphasise this coarse causal role equivalence, aiming to shift attention away from important but ultimately distracting considerations as to whether something adequately fulfils a causal role or not. I also set out an account, via the Extended Digestion analogy, which sought to establish that cognitive extension is just one example amongst many of process extension. In doing so I sought to make the Extended Mind Thesis, and radical cognitive extension, more plausible, but also to shift the onus onto critics who reject the view to make a case for cognitive exceptionalism. Finally, while I rejected Clark and Chalmers Availability and Portability Criteria which are intended to limit what cases count as cognitive extension, I provided a modest starting point to answer an alternative but potentially pressing question which arises if we accept radical cognitive extension: “whose cognition is it?”. My answer was that external cognition belongs to those who are appropriately coupled with it and have meaningful control over it in order to support them in their goals.

8.1 Implications

Accepting the extended mind and embracing cognitive bloat gives us a different view of cognition and of our minds. It moves away from physical boundaries being the boundary of processes. Recognising this is important, because it helps us to recognise the role that external entities play in our cognition. This is not just a matter of the external resources being useful; they can be actively incorporated into our cognitive processes. We adapt to the presence of these external cognitive resources and become skilled users of them, and we store information, beliefs, and even cognitive skills in them. Our minds, then, are not just in our heads (even if they are mainly in our heads),

but may also be in some of our things, some of our surroundings, and even in each other's brains.

If external entities can be so important for our cognition, we may need to grant them a greater importance. Afterall, a theft of Otto's notebook is not just a theft of some paper with writing on it. It is a theft of some of Otto's beliefs, perhaps some of his hopes and dreams. But this is not a case particular to Otto. It is true of all of us, to some extent or other. Parts of our cognition are out there, invested in various things and practices and people. One purpose of the tale about the brick used as a doorstop is to emphasise that cognition is not all so rarefied as we might think and is more commonplace outside of our heads than we might think. The tale of Chompers and his extended digestion draws us to the idea that processes are not just contained within the boundaries of particular objects. And the idea that what makes something our own cognition depends upon our control over it highlights the importance of having control over the external pieces of our minds, and to protect them from the control and influence of others.

8.1.1 You Have a Piece of My Mind

If I tell you to imagine a pink elephant, in particular a pink elephant wearing a blue party hat, I will likely cause you to imagine a pink elephant whether you want to or not. By doing so, I have potentially usurped a small degree of the control you hold over your mind. I have made you imagine something I wanted you to imagine, perhaps even against your will.

In making you think of a pink elephant, I have temporarily reduced the control you have over your own thoughts. But I have not made the thinking about the elephant going on in your brain part of *my* cognition, even though I caused it. This is because it is not participating as part of my cognitive processes.

If, however, I asked you to remember for me my belief that "cheese does not make for a good doorstop", and you commit to doing so, and remind me of this belief when I next ask, then your brain is involved in my cognition, and is part of my cognition. It is also part of your cognition as well. We are sharing part of your brain. Now, in this case the control I have over part of your brain, and the belief stored in it (including access to it) is at a level of control that you have granted me. It is not direct control: it is a limited level of control offered as a service or favour to me, which you could withhold at any time. But while you are willing and able to remember this belief for me and to tell me of it when I next ask, you have a piece of my mind.

9 References

Adams, F. (2019) 'The Elusive Extended Mind: Extended Information Processing Doesn't Equal Extended Mind', in Colombo, M., Irvine, E., and Stapleton, M. (eds.) *Andy Clark and His Critics*. Oxford University Press. USA.

Adams, F. and Aizawa (2001) 'The Bounds of Cognition', *Philosophical Psychology*, Vol.14, No.1, pp.43-64.

Adams, F. and Aizawa, K. (2010) 'Defending the Bounds of Cognition', in Menary, R. (ed.) *The Extended Mind*. MIT Press.

Aizawa, K. (2010) 'The Coupling-Constitution Fallacy Revisited', *Cognitive Systems Research*, Vol.11, Issue 4, pp.332-342.

Anderson, J. (1954) 'Studies on the Cardiac Stomach of the Starfish, Asterias Forbesi', *The Biological Bulletin,* Vol. 107, No. 2.

Bakker, R. (2009) *The Judging Eye*. Orbit.

Banks, I. (2010) *Surface Detail*. Orbit.

Bayne, T. *Philosophy of Mind*. Routledge. New York.

Block, N. (1978) 'Troubles with Functionalism', *Minnesota Studies in the Philosophy of Science*, 9, pp.261-325.

Block, N. (2007) 'What is Functionalism?', *Consciousness, Function, and Representation*, pp.27-44. MIT Press.

Cardinali, L., Zanini, A., Yanofsky, R., Roy, A., Vignemont, F., Culham, J., and Farnè, A. (2021) 'The Toolish Hand Illusion: Embodiment of a Tool Based on Similarity with the Hand', *Scientific Reports (Nature Publishing Group)*, Vol.11, Issue, 1.

Chalmers, D. (2010) 'The Singularity: A Philosophical Analysis'. *Journal of Consciousness Studies*, 17, (9-10), pp. 7-65.

Chalmers, D. (2019) 'Extended Cognition and Consciousness', in Colombo, M., Irvine, E., and Stapleton, M. (eds.) *Andy Clark and His Critics*. Oxford University Press. USA.

Chemro, A. (2001) 'Dynamical Explanation and Mental Representations' in *TRENDS in Cognitive Sciences*, Vol.5, No.4, pp.141-142.

Clark, A. and Chalmers, D. (1998) 'The Extended Mind'. *Analysis*, Vol. 58, No. 1, pp. 7-19.

Clark A. (2008). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford: Oxford University Press.

Clark, A. (2010) '*Memento's* Revenge', in Menary, R. (ed.) *The Extended Mind*. MIT Press.

Clark, A. (2010b) 'Coupling, Constitution, and the Cognitive Kind: A Reply to Adams and Aizawa', in Menary, R. (ed.) *The Extended Mind*. MIT Press.

Clark, A. (2019) 'Replies to Critics', in Colombo, M., Irvine, E., and Stapleton, M. (eds.) *Andy Clark and His Critics*. Oxford University Press. USA.

Couch, M. (2025) 'Causal Role Theories of Functional Explanation', *Internet Encyclopaedia of Philosophy*. Accessed 30/05/2025, available at: https://iep.utm.edu/func-exp/

Dennett. D. (1989) 'The Origin of Selves', *Cogito*, 3, pp.163-173. Accessed 15/04/2021, available at: http://cogprints.org/257/1/originss.htm.

Dennett, D. (2009) 'Where am I?', in Schneider, S. (ed.) *Science Fiction and Philosophy: From Time Travel to Superintelligence*. Wiley-Blackwell. UK.

Farkas, K. (2012) Two versions of the extended mind thesis. *Philosophia*, 40(3), pp. 435-447.

Foster, J. (2009) *Memory: A Very Short Introduction*. Oxford University Press.

Gertler, B. (2007) 'Overextending the Mind?', in Gertler, B. and Shapiro, L. (eds.) *Arguing about the Mind*. Routledge.

Kirsh, D. and Maglio, P. (1992) 'Some Epistemic Benefits of Action: Tetris, a Case Study', in *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*. Morgan Kaufmann, New York.

Kirsh, D. and Maglio, P. (1994) 'On Distinguishing Epistemic from Pragmatic Action', *Cognitive Science*, Vol.18, No.4, pp.513-549.

Levin, J. (2023) 'Functionalism', in Zalta, E. (Ed.) *The Stanford Encyclopaedia of Philosophy* (Summer 2023 Edition). Accessed 09/05/2025, available at: https://plato.stanford.edu/entries/functionalism/.

Knobe, J. and Prinz, J. (2007) 'Intuitions About Consciousness: Experimental Studies' in Chalmers, D. (Ed.) *Philosophy of Mind – Classical and Contemporary Readings*, 2nd Edition, pp.657-668. Oxford University Press.

Kumar, S. and Sahayaraj, K. (2012) 'Gross Morphology and Histology of Head and Salivary Apparatus of the Predatory Bug, Rhynocoris Marginatus', *Journal of Insect Science*, Vol.12, Article 19, pp.1-12.

Meadows, D. (2008) *Thinking in Systems*. Chelsea Green Publishing.

Menary, R. (2010) 'Introduction: The Extended Mind in Focus', in Menary. R. (ed.) *The Extended Mind*. MIT Press.

Morgan, R. (2002) *Altered Carbon*. Victor Gollancz Ltd.

Putnam, H. (1965) 'Brains and Behaviour', in Chalmers, D. (Ed.) *Philosophy of Mind – Classical and Contemporary Readings*, 2nd Edition, pp.61-70. Oxford University Press.

Putnam, H. (1973) 'The Nature of Mental States' in Chalmers, D. (Ed.) *Philosophy of Mind – Classical and Contemporary Readings*, 2nd Edition, pp.79-85. Oxford University Press.

Ross, D. and Ladyman, J. (2010) 'The Alleged Coupling-Constitution Fallacy and the Mature Sciences', in Menary, R. (ed.) *The Extended Mind*. MIT Press.

Rozin, P. (1987) 'A Perspective on Disgust', *Psychological Review*, Vol. 94(1), pp.23-41.

Sahayaraj, K., Kanna, A., & Kumar, S. (2010). Gross Morphology of Feeding Canal, Salivary Apparatus and Digestive Enzymes of Salivary Gland of Catamirus Brevipennis (Servile) (Hemiptera: Reduviidae). *Journal of the Entomological Research Society*, 12(2), pp. 37-50. Accessed 01/08/2019, available at: http://www.entomol.org/journal/index.php/JERS/article/view/176.

Sprevak, M. (2009) 'Extended Cognition and Functionalism', *Journal of Philosophy*, 106, pp.503-527.

Sprevak, M. (2020) 'Extended Cognition', in Crane, T. (ed.) *The Routledge Encyclopaedia of Philosophy Online*. Routledge: London. Accessed 10/05/2025, available at: https://marksprevak.com/publications/extended-cognition/

Star Trek (1989) 'Q Who?', *Star Trek The Next Generation*, Season 2, Episode 16.

Travers, J. (2019) 'How to chew your food properly', *The Guardian*. Accessed 08/08/2019, available at: https://www.theguardian.com/lifeandstyle/2019/jul/14/how-to-chew-your-food-properly.

Wheeler, M. (2010) 'In Defence of Extended Functionalism', in Menary, R. (ed.) *The Extended Mind*. MIT Press.

Vallar, G. and Ronchi, R. (2009) 'Somatoparaphrenia: a body delusion. A review of the neuropsychological literature', *Experimental Brain Research*, 192, pp.533-551.

Zelazny, R. (1967) *Lord of Light*. Reprint, Millennium 2004.